2011
2012

FACULTY OF BUSINESS ECONOMICS

Master of Management: Management Information Systems

Masterproef

Developing a process mining workflow based on the KDD framework

Promotor :
Prof. dr. Benoit DEPAIRE

Raf Geuns

Master Thesis nominated to obtain the degree of Master of Management , specialization Management Information Systems

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt

universiteit hasselt
KNOWLEDGE IN ACTION

2011
2012

FACULTY OF BUSINESS ECONOMICS

Master of Management: Management Information Systems

Masterproef

Developing a process mining workflow based on the KDD framework

Promotor :
Prof. dr. Benoit DEPAIRE

Raf Geuns

Master Thesis nominated to obtain the degree of Master of Management , specialization Management Information Systems

universiteit hasselt
KNOWLEDGE IN ACTION

PREFACE

Writing a master thesis is an important and obligated subject for gaining the title Master of Management. As many know, a lot of time is spent in writing a master thesis, but this gives an instructive experience.

First I would like to thank my promoter prof. dr. Benoît Depaire for all of his guidance. Also I would like to thank the interviewees for taking the time and effort for providing the information for completing this master thesis.

Finally I would like to thank my parents for giving me the opportunity to complete this master year. Also I would like to thank my friend in supporting me throughout the year.

Raf Geuns

SUMMARY

In the beginning of 2012, the IEEE Task force on Process Mining created a list of eleven challenging aspects concerning process mining. A couple of these challenges state that non-experts find it difficult to perform process mining and analyze process mining results. The goal of this master thesis is try to present an effective methodology for gaining knowledge from business processes with the use of process mining.

This master thesis contains seven chapters. In the first chapter, the problem definition is defined. Based on this, a research question is formulated. Next to this, three sub questions are stated that try to give an answer on the research question. This chapter also contains the research methodology that was used to perform this master thesis.

In the second chapter preliminaries of process mining are presented. First, different process languages, such as Petri nets and Business Process Modeling Notation, are presented and explained. Finally, the most common control-flow constructs are clarified.

The two dimensions of process mining are discussed in the third chapter. Based on these two dimensions, fourteen types of process mining analyses can be distinguished. For every analysis, the goals and history are explained. Also many process mining tools, methods and techniques with their assumptions and limitations are presented. After this, a historical evolution of process mining problems is mentioned. As a last part, the current process mining challenges are explained.

Process mining is derived from data mining, which is a part of Knowledge Discovery in Databases (KDD). First, different process models from KDD are presented. After this, these are further analyzed.

In the fifth chapter, the steps from the KDD process models are examined if they are possible to use in a process mining framework. Also an existing process diagnostics methodology is mentioned

and explained. Based on these findings, a process mining framework is proposed in the last subsection. The differences and resemblances between the KDD process models and the new process mining framework are discussed in the sixth chapter.

The seventh and final chapter contains the conclusions and recommendations. In the conclusions, an answer to the research questions and the sub question is given. Finally, additional recommendations are formulated for possible future research.

TABLE OF CONTENTS

PREFACE 1

SUMMARY 3

TABLE OF CONTENTS 5

TABLE OF FIGURES 9

TABLE OF TABLES 11

CHAPTER 1: Research plan 13

1.1 Problem definition 13

1.2 Research question and sub questions 17

1.2.1 Research question 17

1.2.2 Sub questions 18

1.3 Research methodology 19

CHAPTER 2: Preliminaries 21

2.1 Process languages 21

2.1.1 Petri nets 21

2.1.2 Workflow nets 23

2.1.3 Business Process Modeling Notation (BPMN) 23

2.1.4 Event-Driven Process Chains (EPC) 25

2.2 Control-flow constructs 26

CHAPTER 3: Process Mining 29

3.1 Dimensions of process mining 29

3.2 Process mining analyses 31

3.2.1 Control-flow discovery 31

3.2.2 Control-flow conformance 36

3.2.3 Control-flow enhancement 37

3.2.4 Organizational discovery 38

3.2.5 Organizational conformance 39

3.2.6 Organizational enhancement 40

3.2.7 Case discovery 40
3.2.8 Case conformance 41
3.2.9 Case enhancement 42
3.2.10 Time discovery 42
3.2.11 Time conformance 43
3.2.12 Time enhancement 43
3.3 Process mining problems 44
3.4 Process mining challenges 45
CHAPTER 4: Data mining and knowledge discovery in databases 49
4.1 KDD process models 49
4.2 Analysis of the KDD process models 51
4.2.1 Step 1: application domain understanding 51
4.2.2 Step 2: data understanding 55
4.2.3 Step 3: data preparation and identification of data mining technology 56
4.3.4 Step 4: data mining 56
4.3.5 Step 5: evaluation 56
4.3.6 Step 6: knowledge consolidation and deployment 57
CHAPTER 5: Process mining framework 59
5.1 Linking KDD steps to process mining 61
5.1.1 Step 1: application domain understanding 61
5.1.2 Step 2: data understanding 62
5.1.3 Step 3: data preparation and identification of data mining technology 62
5.1.4 Step 4: data mining 63
5.1.5 Step 5: evaluation 63
5.1.6 Step 6: knowledge consolidation and deployment 63
5.2 Process diagnostics 64
5.3 Development of a process mining framework 66
5.3.1 Step 1: goal definition 66
5.3.2 Step 2: search and preparation of data 67

5.3.3 Step 3: creation of event logs and identification of process mining technology 68
5.3.4 Step 4: process mining 69
5.3.5 Step 5: evaluation 69
5.3.6 Step 6: knowledge consolidation and deployment 70
CHAPTER 6: Discussion 73
CHAPTER 7: Conclusions and recommendations 75
7.1 Conclusions 75
7.2 Recommendations 76
REFERENCES 77
APPENDICES 87
A Tables 87

TABLE OF FIGURES

Figure 1.1: Moore's Law 13
Figure 1.2: Internet Users in the World: Growth 1995-2010 14
Figure 1.3: Exponential Data Warehouse Growth: size in terabytes of user data 15
Figure 2.1: A marked Petri net 22
Figure 2.2: Process model using BPMN notation 24
Figure 2.3: BPMN notation 24
Figure 2.4: Process model using EPC notation 25
Figure 2.5: EPC notation 25
Figure 2.6: A Petri net with the common control-flow constructs 26
Figure 2.7: Possible parallelism structures 27
Figure 2.8: Possible choice structures 27
Figure 2.9: Possible loop structures 28
Figure 2.10: Possible non-free-choice structures 28
Figure 3.1: Three types of process mining 29
Figure 3.2: The process model extended with additional perspectives 30
Figure 3.3: Papers dealing with process mining problems 45
Figure 3.4: Process mining problems 45
Figure 4.1: Steps of the KDD process 49
Figure 5.1: The phases of the process diagnostics methodology 65
Figure 5.2: Process mining flowchart 66

TABLE OF TABLES

Table 1.1: An event log 16
Table 2.2: Overview of the event log in table 2.1 21
Table 3.1: Type/Perspective-matrix 31
Table 3.3: Control-flow discovery algorithms, tools and methods 32
Table 3.4: Comparison of Petri net discovery algorithms 35
Table 3.5: Control-flow conformance algorithms, tools and methods 36
Table 3.6: Control-flow enhancement algorithms, tools and methods 38
Table 3.7: Organizational discovery algorithms, tools and methods 38
Table 3.8: Organizational conformance algorithms, tools and methods 39
Table 3.9: Organizational enhancement algorithms, tools and methods 40
Table 3.10: Case discovery algorithms, tools and methods 41
Table 3.11: Time discovery algorithms, tools and methods 43
Table 3.12: Time enhancement algorithms, tools and methods 44
Table 4.1: Comparison of the major existing KDD models 50
Table 4.2: Detailed description of individual steps of the major existing KDD models 52
Table 5.1: Summary of the KDD models 59
Table 5.2: Goal definition 67
Table 5.3: Search and preparation of data 68
Table 5.4: Creation of event logs and identification of process mining technology 68
Table 5.5: Process mining 69
Table 5.6: Evaluation 70
Table 5.7: Knowledge consolidation and deployment 70
Table 5.8: Process mining framework 71

CHAPTER 1: Research plan

1.1 Problem definition

In the world of technology, there has been a major explosion on different aspects since the 1950s. One example is the number of components in integrated circuits. Moore's law (figure 1.1) stated that the number of components in integrated circuits would double every year since 1965 (*Moore's Law*, 2011). When we look back, we can conclude that this was almost the case, although at a slightly slower pace.

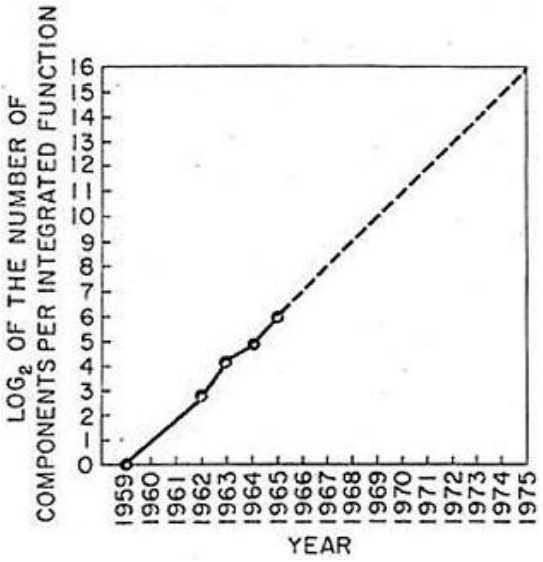

Figure 1.1: Moore's Law (*Moore's Law Graph*, 2011)

Another example is the growth of data. It is very difficult to put an exact number on the amount of data that has been created thanks to the internet. Figure 1.2 gives an overview of the growing number of internet users since 1995 until 2010. In 1995, 16 million users, or 0.4% of the world population was connected to the internet. In 2010, the amount of internet users were expected to reach to 2,110 million or 30.4% of the world population (*Internet Growth Statistics*, 2011). To give an example, Facebook reached 1 million active users in December 2004. By July 2011 this increased to an astonishing number of 750 million active users (*Timeline*, 2011). Each of these

users can make a profile and upload photos or videos on their profile. These photos or videos are nothing more than data that can be found on the servers of Facebook. The growing amount of Facebook and internet users in general, will cause an increasing amount of data.

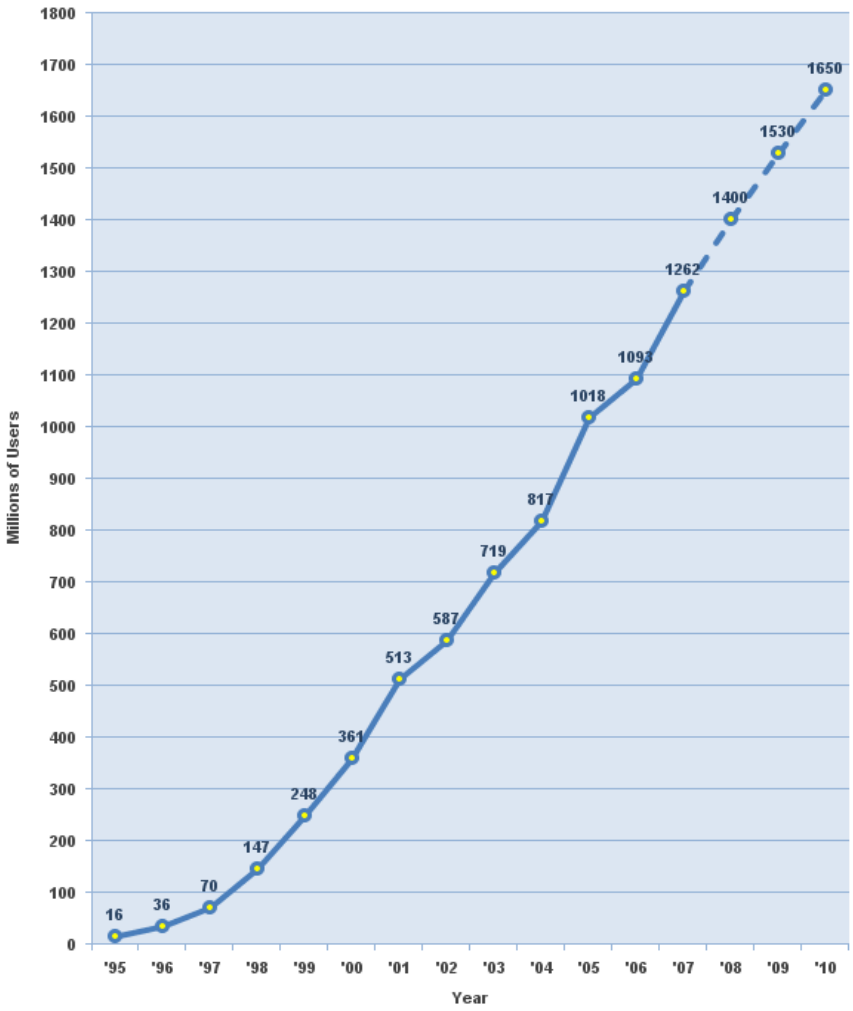

Figure 1.2: Internet Users in the World: Growth 1995-2010 (*Internet Growth Statistics*, 2011)

With the growth of data, there also has to be a growth in the storage capacity of a data warehouse. Figure 1.3 states that the actual growth of the data warehouse is larger than Moore's Law growth rate predicted. To analyze these huge amount of data, the Apache Software Foundation designed the open-source project Hadoop. It can work with up to petabytes of data. Hadoop used a technique called MapReduce: high-performance parallel data processing. In addition, Hadoop connects different servers with each other so that a cluster is created. (*What is Hadoop?*, 2011).

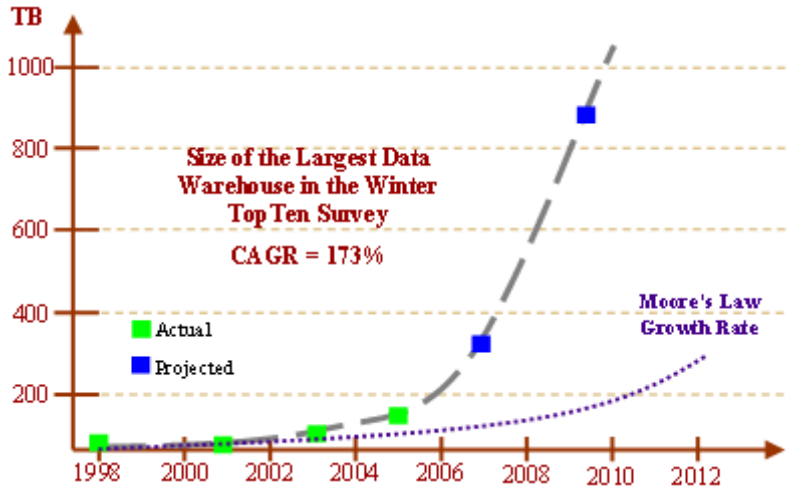

Figure 1.3: Exponential Data Warehouse Growth: size in terabytes of user data (Winter, 2008)

There is not only a growth of data, but also a growing role and importance of information systems to control these data. These information systems know a historical development. In the 60s, data was used to forecast the future need of goods. In the 70s, companies started using systems for Materials Requirement Planning (MRP), which integrates production and planning. In the 80s, MRP was supplemented with capacity planning (MRP-II). The MRP-II-systems were combined with Manufacturing Execution Systems (MES) in the 90s. This gave the opportunity to tune in on the customer's needs. At the end of the 90s, ERP-systems were developed to integrate information about the suppliers, the manufacturing and the customers in the supply chain (Summer, 2007). Other examples of information systems are WorkFlow Management (WFM), Customer Relationship Management (CRM), Supply Chain Management (SCM) and Product Data Management (PDM). They all belong to the category of Process-Aware Information Systems (PAISs), which means systems are more focused on processes instead of data (Weijters, van der Aalst & Alves de Meideros, 2006).

All of these PAISs record information about business processes in the form of event logs (van der Aalst et al., 2008). When a process is executed, is refers to a case. A case consists of different events. When an event is executed, it refers to an activity, which is a well-defined step in the process. An activity can have a performer/originator, the person who executes or initiated the activity, and a time stamp (Weijters et al., 2006). Table 1.1 gives an example of an event log

which holds 5 different cases, numbered from 1 to 5. Each of these cases have different activities, which have a letter going from A to E. We can also remark that there are 5 different originators. Here we can see that de cases chronologically ordered by their time stamp.

Table 1.1: An event log (Weijters et al., 2006)

case id	activity id	originator	time stamp
case 1	activity A	John	9-3-2004:15.01
case 2	activity A	John	9-3-2004:15.12
case 3	activity A	Sue	9-3-2004:16.03
case 3	activity B	Carol	9-3-2004:16.07
case 1	activity B	Mike	9-3-2004:18.25
case 1	activity C	John	10-3-2004:9.23
case 2	activity C	Mike	10-3-2004:10.34
case 4	activity A	Sue	10-3-2004:10.35
case 2	activity B	John	10-3-2004:12.34
case 2	activity D	Pete	10-3-2004:12.50
case 5	activity A	Sue	10-3-2004:13.05
case 4	activity C	Carol	11-3-2004:10.12
case 1	activity D	Pete	11-3-2004:10.14
case 3	activity C	Sue	11-3-2004:10.44
case 3	activity D	Pete	11-3-2004:11.03
case 4	activity B	Sue	11-3-2004:11.18
case 5	activity E	Clare	11-3-2004:12.22
case 5	activity D	Clare	11-3-2004:14.34
case 4	activity D	Pete	11-3-2004:15.56

Nowadays, business processes are an essential element of the organizational structure. They can manage and coordinate the activities within a company (Bibiano, Mayol, & Pastor, 2007). Business processes, compared to technology, cannot easily be copied by other companies, so this is how enterprises create sustainable value nowadays. Peppard and Ward (2005) state that building relationships with customers is a process. Apple has created its business process in such a way it mostly focuses on user experience (Rundle, 2010). This sort of experience is something very difficult to copy by competitors, so they must focus on some other aspects to compete with.

The main problem about business processes is that most managers and board members don't know them (Champy, 1995). This is where process mining comes in handy. With the help of the event logs generated by PAISs, process mining can discover business processes using algorithms. Some examples of algorithms are the Alpha (α) (van der Aalst, Weijters & Maruster, 2004), Alpha^{++} (α^{++}) (Wen, Wang & Sun, 2005), Heuristics Miner (Weijters et al., 2006), Genetic Miner (de Medeiros,

Weijters & van der Aalst, 2007) and Region Miner (Bergenthum, Desel, Lorenz & Mauser, 2007) algorithms (Weber, 2009). Different algorithms have been developed for different reasons, for example, for creating Petri nets, for discovering relationships, etc. Many different tools and techniques have been developed over the last couple of years. The problem of most of these techniques is that they make assumptions which do not hold in practical situations (van der Aalst, 2007).

When conducting process mining, it is useful to use a framework such as ProM. ProM is based on open-source and is used for the implementation of process mining tools and techniques (*ProM*, 2009). Currently, the framework provides more than 230 plug-ins. These can be subdivided into 5 types of plug-ins (van der Aalst et al., 2010):

- Mining plug-ins: use mining algorithms to construct a Petri net;
- Export plug-ins: provide a "save as" functionality for objects like graphs;
- Import plug-ins: implement an "open" functionality for exported objects;
- Analysis plug-ins: implement some property analysis on some mining results;
- Conversion plug-ins: implement conversions between different data formats.

As stated above, many tools and techniques exist. This is where the main problem of process mining can be found: how does one begin process mining on a structured manner?

1.2 Research question and sub questions

1.2.1 Research question

As stated in the problem definition, process mining has a lot to offer. Despite this, it also has quite some problems. In 2012, the IEEE Task Force of Process Mining published a manifesto in which they stated several principles and challenges (van der Aalst et al., 2012). Instead of only focusing on problems like hidden tasks, noise, duplicate tasks, etc. (van der Aalst & Weijters, 2004), they also proposed challenges in a more global way. Examples are cross-organizational mining,

providing operational support, combining process mining with other types of analysis, improving usability and understandability for non-experts, etc.

In the problem definition, we have mentioned that there are a lot of different process mining algorithms and tools with different purposes. But since there are so many, how can one compare them? Rozinat, Alves de Medeiros, Günther, Weijters & van der Aalst (2007a) state that although "process mining reached a certain level of maturity … a common framework to evaluate process mining results is still lacking" (p.1). This comes with a fact that researchers can't compare the performance of different process mining algorithms. Since no framework exists, it can be hard to find the right algorithm or tool to perform a certain process mining analysis. This is why the following research question is chosen:

"What is an effective methodology to gain knowledge from business processes with the use of process mining techniques?"

1.2.2 Sub questions

To answer the research question, we first must ask some sub questions in function of the research question. As stated in the problem definition, a wide range of techniques for process mining have been developed throughout the years. It can come in handy to research and discuss the most popular ones, or if it is possible to create clusters for similar techniques.

The first following sub question is:

"What different types of process mining techniques do exist?"

Before an effective methodology can be made, we first have to conduct research and try to find the different methodologies used in process mining. The advantages and disadvantages will be mapped. Based on this, the different methodologies will be compared with each other.

The second following sub question is:

"What methodologies do exist for knowledge discovery in databases?"

As stated before, process mining can be linked with data mining, which is a part of knowledge discovery in databases. Therefore it can be useful to investigate whether the methodologies for knowledge discovery in databases can also be used for process mining.

The last sub question is:

"How can the methodologies for knowledge discovery in databases be used for process mining?"

1.3 Research methodology

The articles were found using Google Scholar and scientific databases. Hasselt University provides a large list of databases the university has subscribed to. Some examples of these databases are EBSCOhost, AtoZ, Academic Search Elite, Business Source Premier, etc. Google Scholar is used to research with articles are cited most and by which authors. Google Scholar also provides an easy access to most articles. When access was limited, the title of the magazine was looked up in AtoZ, which gives an overview of all the scientific journals Hasselt University has subscribed to. Articles were also found by studying the reference lists of the read articles. In Google Scholar, there was also the possibility to check for authors and articles who have referred to a read article.

Articles concerning process mining were found using following terms: process mining, process discovery, algorithm, framework, social network, evaluation. Most articles were found in scientific magazines as Lecture Notes in Computer Science and Data & Knowledge Engineering. One of the most important articles was the process mining manifesto (van der Aalst et al., 2012), which presented a series of guidelines and challenges. The topic of this thesis has a connection with some

of these challenges. Another important source about process mining, was a book written by van der Aalst (2011), which included a lot of basic information.

Articles concerning knowledge discovery in databases and data mining were found using following terms: knowledge discovery in databases, data mining, process models, framework, guidelines. The most important article was written by Kurgan & Musilek (2006). It gave an overview of the five most used knowledge discovery in databases process models. Only an article concerning the framework proposed by Fayyad, Piatetsky-Shapiro & Smyth (1996b) was found. The other process models were published in books, which could not be accessed.

At the end of the research, a few process mining experts were interviewed. They were asked which series of tasks they performed when conducting process mining. These answers were compared with the knowledge discovery in databases process models, and based on this, a flowchart for process mining was made.

CHAPTER 2: Preliminaries

2.1 Process languages

In this part, different type of languages will be discussed and explained by an example of van der Aalst (2011). Table 2.1 in appendices gives a small part of an event log. Table 2.2 gives a more compact overview. The following letters are linked to an activity:

- a = register request;
- b = examine thoroughly;
- c = examine casually;
- d = check ticket;
- e = decide;
- f = reinitiate;
- g = pay compensation;
- h = reject request.

Table 2.2: Overview of the event log in table 2.1 (van der Aalst, 2011)

Case id	Trace
1	$\langle a, b, d, e, h \rangle$
2	$\langle a, d, c, e, g \rangle$
3	$\langle a, c, d, e, f, b, d, e, g \rangle$
4	$\langle a, d, b, e, h \rangle$
5	$\langle a, c, d, e, f, d, c, e, f, c, d, e, h \rangle$
6	$\langle a, c, d, e, g \rangle$
...	...

2.1.1 Petri nets

One of the best investigated process modeling languages are Petri nets. It is a rather simple graphical notation, but nevertheless many analysis techniques can be used to analyze them. A Petri

net consists of a finite set of places, a finite set of transitions and a set of directed arcs (van der Aalst, 2011). A place can contain a token, which can flow through the network when it is fired. When a place contains a token, it is called a marking (Weber, 2009). A Petri net has a mathematical representation, but this is beyond the scope of this thesis. Figure 2.1 shows a Petri net. We can see that place "start" contains a token. This means that transition "a" is enabled. Enabling only takes place when all the input places of a transition contain a token. Since transition "a" is enabled, it takes the token from place "start" and gives a token to each of its output places, namely "c1" and "c2". This is called the firing rule. Now, transition "a" isn't enabled anymore, but transitions "b", "c" and "d" are. When transition "b" or "c" is fired, place "c3" contains a token. Transition "e" can only be enabled when transition "d" is fired. When we look at place "c5", we see that there is a loop, namely transition "f". This results in a infinite firing sequence starting in place "start" and ending in place "end".

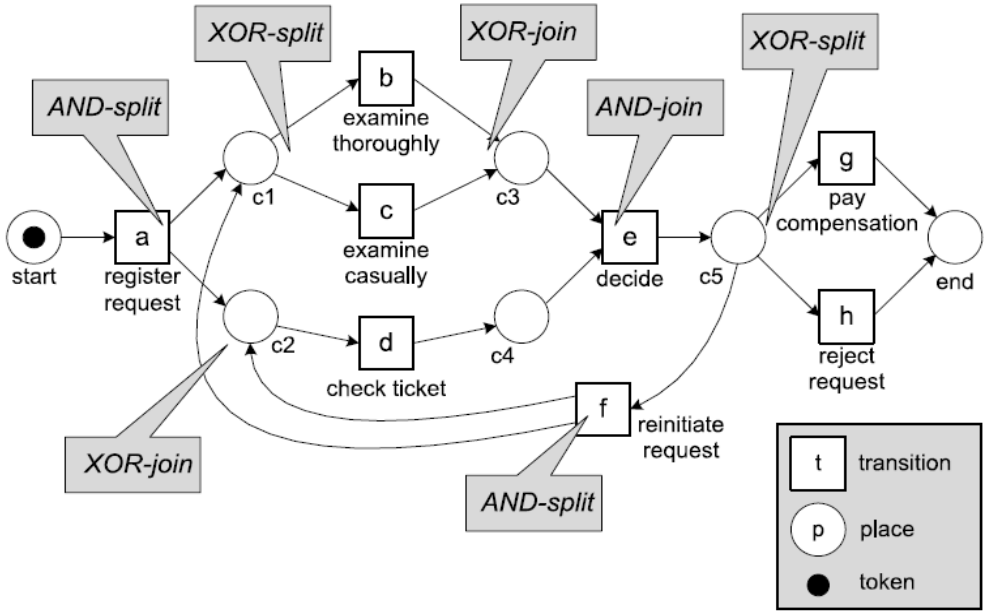

Figure 2.1: A marked Petri net (van der Aalst, 2011)

In this figure, we can also see splits and joins. An AND-split means that all of the outputs must be chosen. When we look at the example, firing transition "a" leads to a token in places "c1" and "c2". A XOR-split means that exactly one of its outputs must be chosen. For this example, when place

"c1" contains a token, transitions "b" and "c" are enabled. Since it is a XOR-split, only one of these transitions can be chosen. The chosen transition always leads to place "c3", which makes it a XOR-join. Transition "e" is an AND-join, because two places are joined here, namely places "c3" and "c4". Place "c4" can contain a token when transition "d" is fired, which is enabled when place "c2" contains a token.

2.1.2 Workflow nets

A Workflow net is a subclass of Petri nets. Workflow nets have a single start (source) and end (sink) place. A Petri net can be considered as a Workflow net if and only if (van der Aalst, 2011):

- The set of places contains an input place *i*;
- The set of places contains an output place *o*;
- There is a directed path between any pair of nodes in the Petri net.

Figure 2.1 represents a Workflow net, since place "start" is a unique source and place "end" is a unique sink. But we must keep in mind that not every Workflow net represents a correct process, since some errors may occur. To have a correct Workflow, we must take following conditions into account (van der Aalst, 2011):

- Safeness: places cannot hold multiple tokens at the same time;
- Proper completion: when the process execution is finished, i.e. the place "end" contains a token, there can't be any token left at the other places;
- Option to complete: for any reachable marking, there must be a path to the place "end";
- Absence of dead parts: there are no dead transitions, which means that for any transition, there is a firing sequence enabling that transition.

2.1.3 Business Process Modeling Notation (BPMN)

BPMN has become a widely used business process modeling language and is supported and standardized by many tools. Figure 2.2 gives an example of a BPMN, using the information from

table 2.1 and 2.2. When we compare the BPMN and Petri net, we can see that an event is quite similar to a place. An event can only have one incoming and outgoing arc, splitting and joining must be done using gateways as shown in figure 2.3. This is a difference with places in a Petri net. In figure 2.1 we can see that the event "end" has two incoming arcs and quite some events have multiple incoming and/or outgoing arcs. The layout of the different splitting and joining of Petri nets and BPMN is different, but the function remains the same (van der Aalst, 2011).

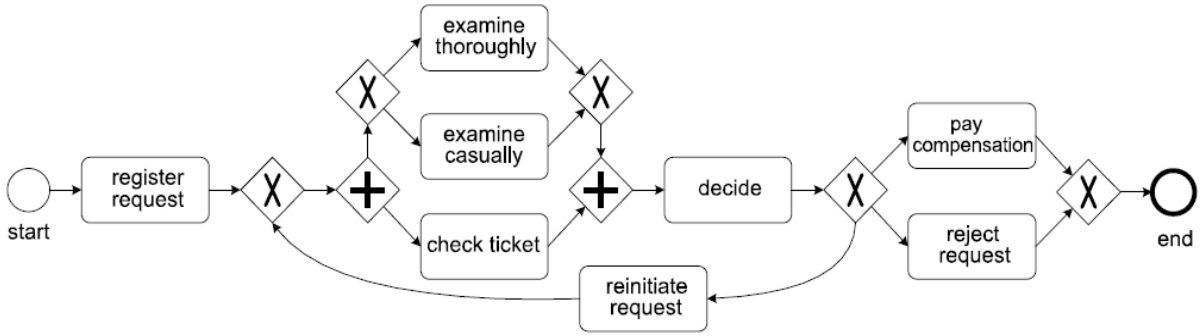

Figure 2.2: Process model using BPMN notation (van der Aalst, 2011)

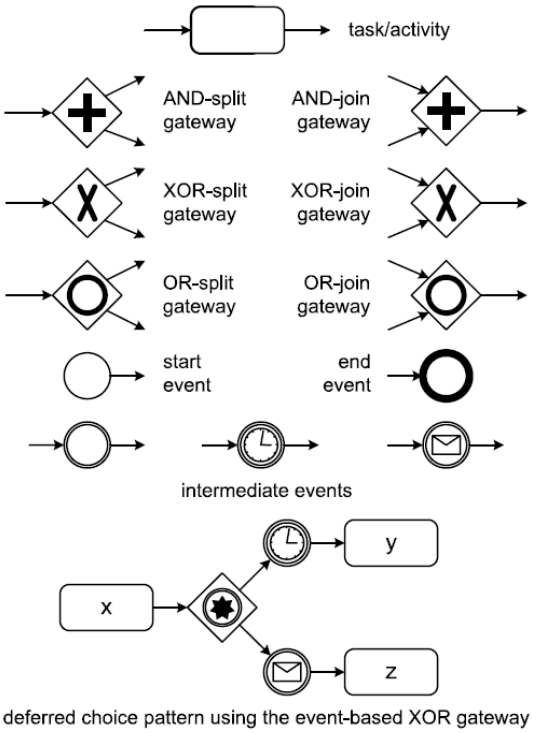

Figure 2.3: BPMN notation (van der Aalst, 2011)

2.1.4 Event-Driven Process Chains (EPC)

When comparing the notations of EPC (see figure 2.5) and BPMN, we can see quite some similarities. Functions and activities have the same meaning. A function can only have one input and one output arc, so the use of joins and splits is required. Also here there are AND, XOR and OR types. There can also be made a distinction between three types of events: start, intermediate and end. This is the same as BPMN. Also, events cannot be connected to events or functions cannot be connected to functions. When we look at figure 2.4, we can see process model using EPC.

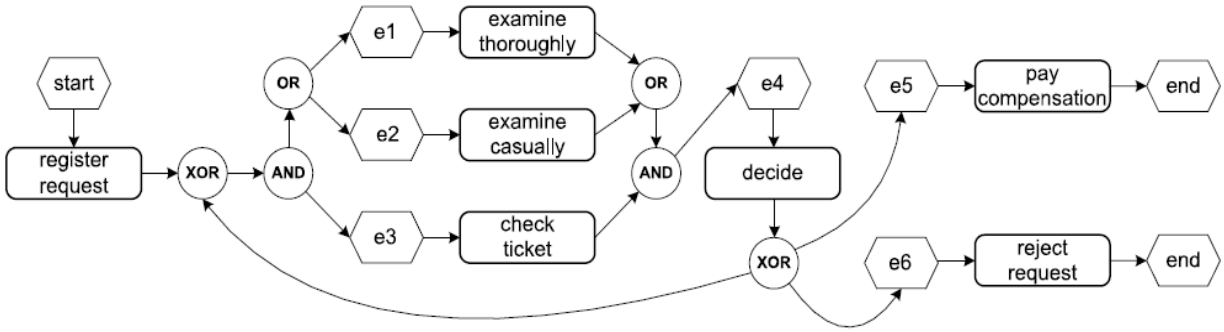

Figure 2.4: Process model using EPC notation (van der Aalst, 2011)

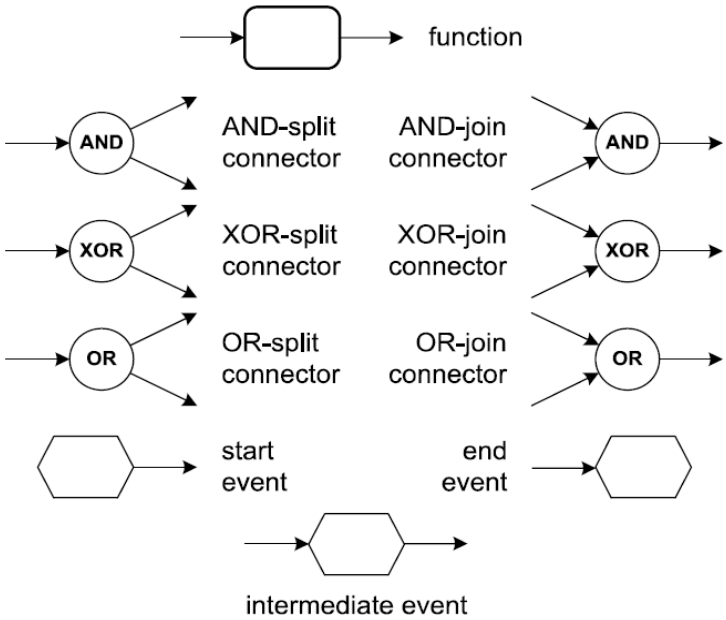

Figure 2.5: EPC notation (van der Aalst, 2011)

2.2 Control-flow constructs

The mined process model is an objective overview of possible flows that were followed by cases in event logs. Since the flow of tasks will be visualized, it is important that process discovery techniques are able to support the correct mining of the common control-flow constructs. These constructs are sequence, parallelism, choices, loops, non-free-choice, invisible tasks and duplicate tasks (van Dongen, Alves de Medeiros & Wen, 2009). A representation of a Petri net with all these constructs in given in figure 2.6.

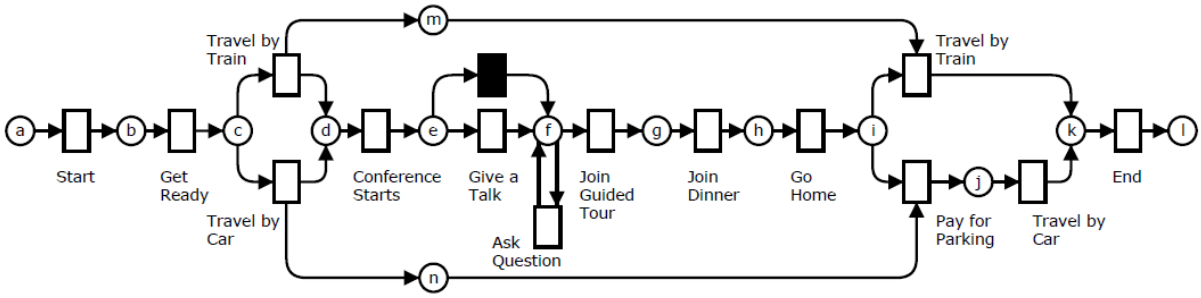

Figure 2.6: A Petri net with the common control-flow constructs (van Dongen et al., 2009)

Sequence is the course of all the tasks within the model. Parallelism occurs when an AND split and join are present in the model. When following an AND split, the token splits up and follows all the paths that leave the AND split. All these tasks are executed in parallel. Figure 2.7 gives some examples of parallelism. Choice occurs when an OR split and join are present in the model. Since only one path can be followed, a choice must be made. Some examples can be found in figure 2.8. Loops take place when an XOR split has one path that leads back to that same split. Figure 2.9 gives examples of loop structures. Non-free-choice means that that choice of paths have an impact on what choices can be made later. Some examples are presented in figure 2.10. Invisible tasks are tasks that have no corresponding log event. Duplicate tasks are multiple tasks that are mapped onto the same kind of event log (Rozinat & van der Aalst, 2008; Weber, 2009).

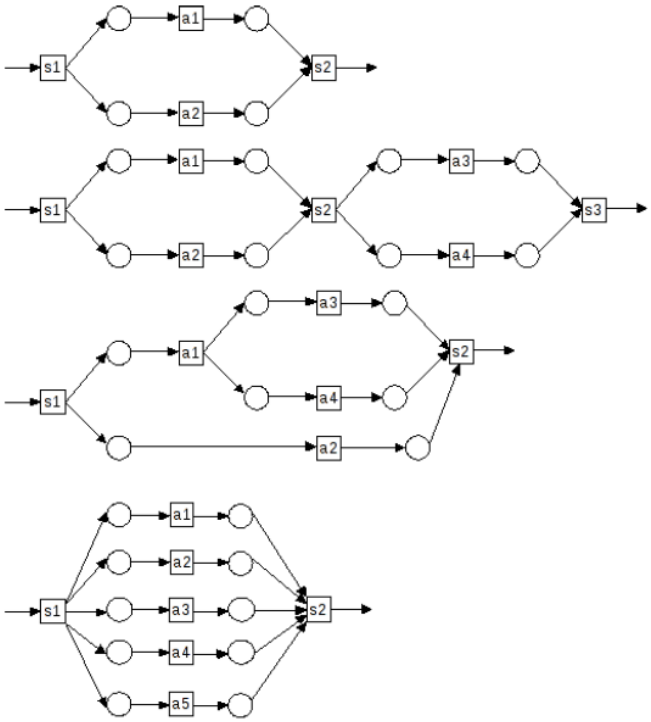

Figure 2.7: Possible parallelism structures (Weber, 2009)

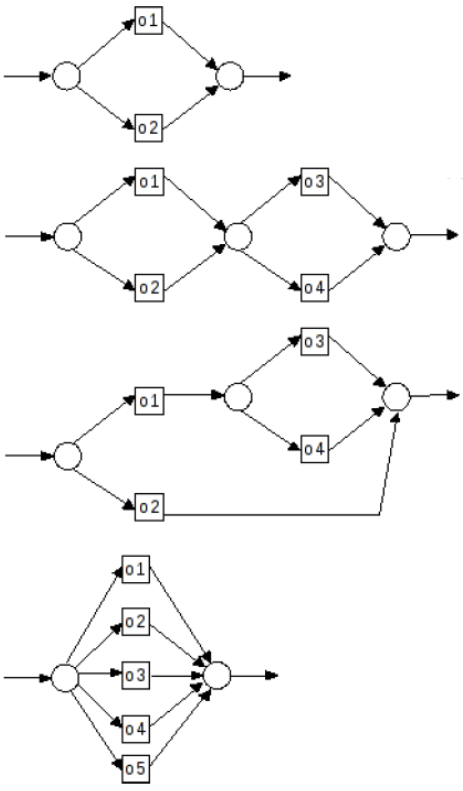

Figure 2.8: Possible choice structures (Weber, 2009)

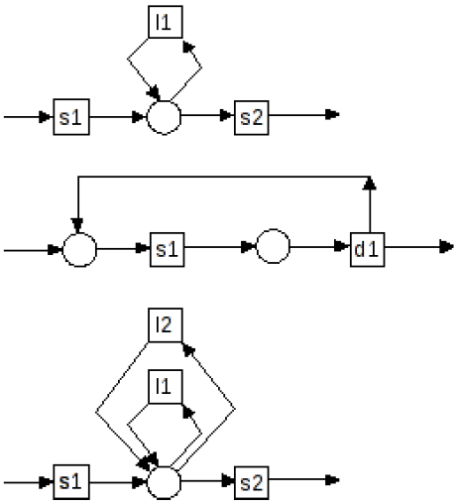

Figure 2.9: Possible loop structures (Weber, 2009)

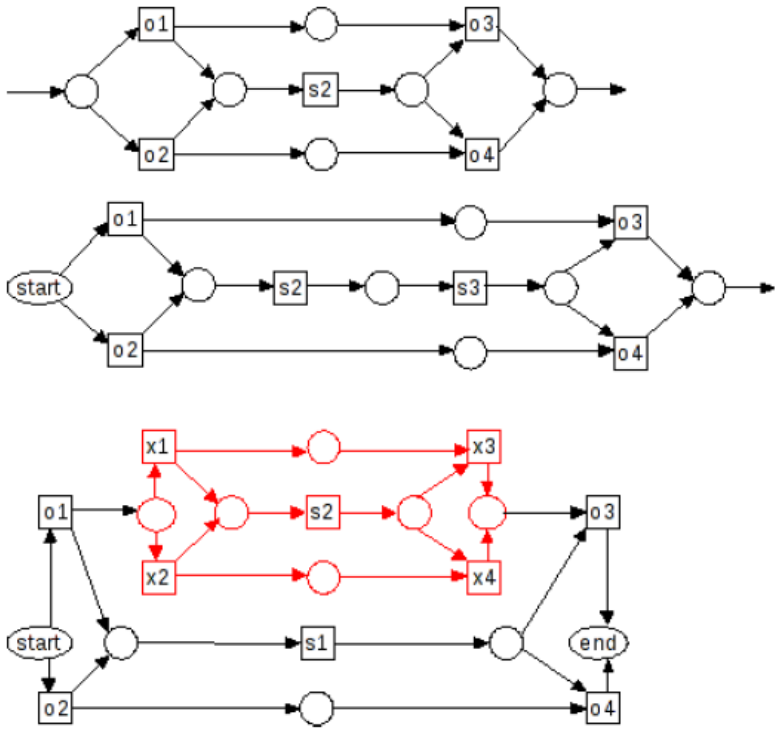

Figure 2.10: Possible non-free-choice structures (Weber, 2009)

CHAPTER 3: Process Mining

As stated in the problem definition, process mining uses data from event logs (which are generated by PAISs) to discover business processes. This chapter will first handle the different dimensions of process mining. Then the different types of process mining analyses will be discussed.

3.1 Dimensions of process mining

Mans, Schonenberg, Song, van der Aalst and Bakker (2008) say that "the idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs" (p. 427). Based on this statement, three types of process mining can be considered: discovery, conformance and extension (see figure 3.1). Discovery means that there is no a-priori model. Conformance means that there is an a-priori model and is used to check if the actual model conforms to the predefined model. Enhancement (or extension) is the last type. There is also an a-priori model and the goal is to enrich the model using information from event logs. It is important to state that process mining is not limited to models. Since event logs consist of data about performers or time, it is possible to evaluate social structures within an organization (van der Aalst et al., 2006).

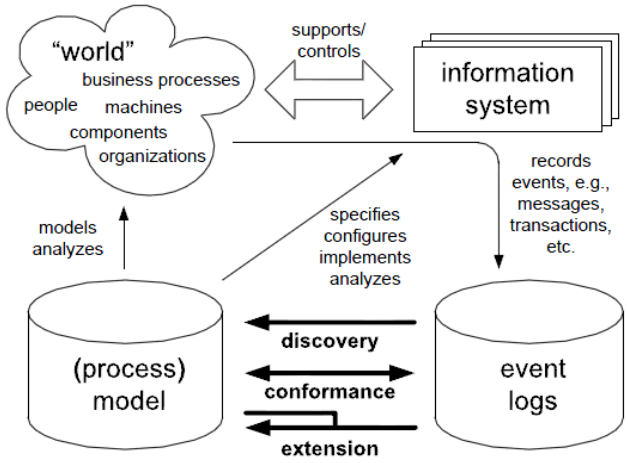

Figure 3.1: Three types of process mining (van der Aalst, Rubin, van Dongen, Kindler and Günther, 2006)

Apart from these three types of process mining, there are also four different kind of perspectives. The first one is the control-flow perspective and focuses on the ordering of activities, which are represented in terms of a Petri net, Workflow, EPC, BPMN, etc. The second perspective is the organizational perspective. It focuses on the organizational structure and resources like people, systems, roles and departments. The third one is the case perspective. This perspective focuses on the details of a case, for example a supplier or the amount of ordered goods. The last perspective is the time perspective and focuses on timing and frequency. This is only possible if the event log contains a time stamp. Using time stamps, one of the advantages of this perspective can be to discover bottlenecks. The different types and perspectives can occur orthogonal (van der Aalst, 2011). Figure 3.2 gives a graphical representation of a process model using all the process mining perspectives.

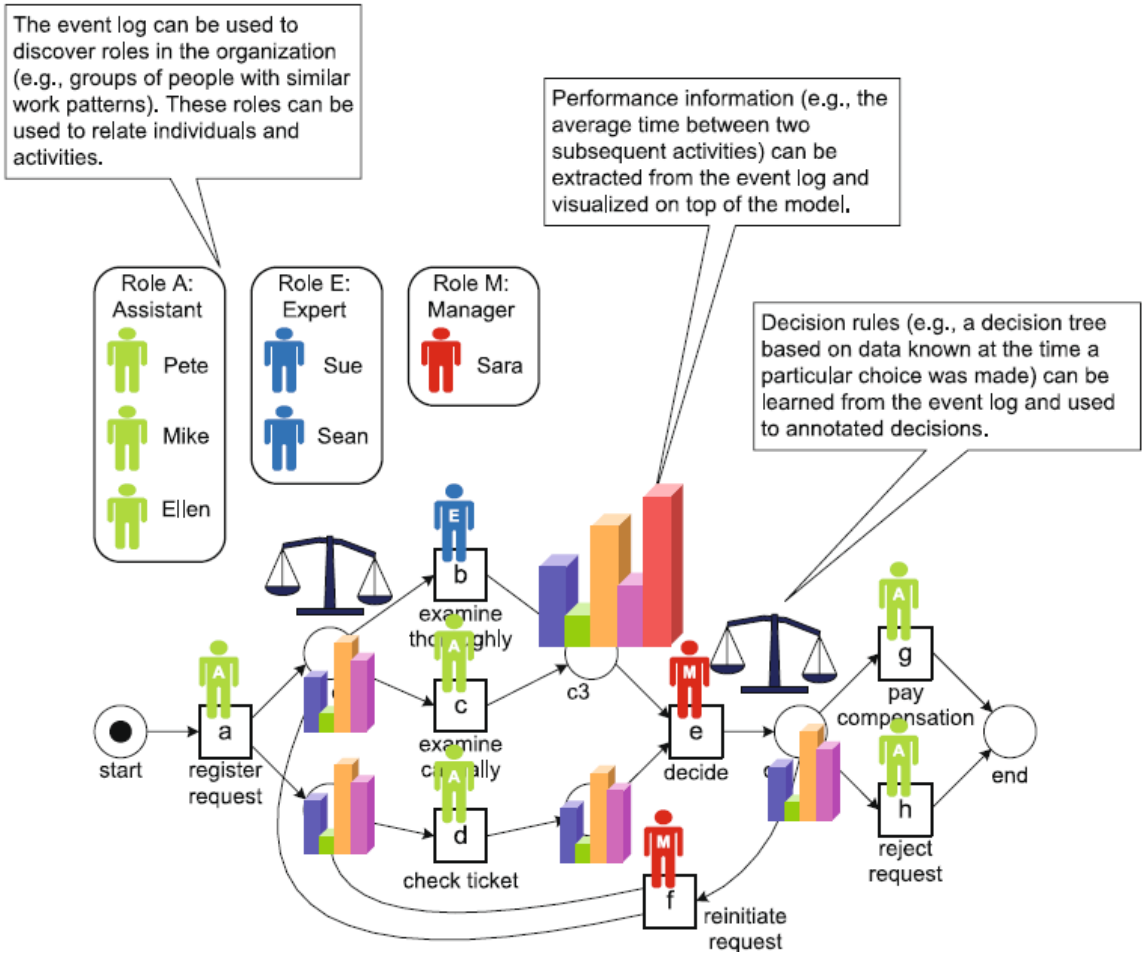

Figure 3.2: The process model extended with additional perspectives (van der Aalst, 2011)

Based on research mentioned above, process mining can be divide into two dimensions. The structure is given in table 3.1. As we can see, the matrix consists fourteen analyses. The different kind of analyses are control-flow discovery, control-flow conformance, control-flow enhancement, organizational discovery, organizational conformance, organizational enhancement, case discovery, case conformance, case enhancement, time discovery, time conformance and time enhancement. Based on the conducted literature study, the matrix (table 3.1) has been filled in with "+" or "0". More "+" indicate more research papers about the analysis in question. As we can see, control-flow discovery is the most discussed analysis. No articles were found about case conformance, case enhancement and time conformance. Table 3.2 in appendices gives an overview of the articles that discuss one or more of the fourteen analyses.

Table 3.1: Type/Perspective-matrix

		TYPES OF PROCESS MINING		
		Discovery	**Conformance**	**Enhancement**
PERSPECTIVE	**Control-flow**	++++	+++	+
	Organization	++	+	+
	Case	++	0	0
	Time	+	0	++

3.2 Process mining analyses

3.2.1 Control-flow discovery

The goals of control-flow discovery is trying to give an overview of the possible paths of activities within an organization. Information is taken from event logs without any a-priori information and then visualized with the help of process languages (Petri nets, Workflow nets, BPMN, EPC, etc.), which are described in the previous chapter.

Table 3.3 gives an overview of the discovered algorithms, tools and methods that are able to discover the control-flow. When we analyze the table, we can see that there have been many different algorithms, tools and methods developed. The roots of process mining lie around mid and the end of the 90's. Agrawal, Gunopulos and Leymann (1998) were one of the first to conduct process mining within the context of workflow management. Cook and Wolf (1998a) used three methods: a neural network (RNet), a algorithmic approach (KTail) and a Markovian approach. They extended the research by proposing specific metrics (entropy, event type counts, periodicity and causality) to discover models (Cook & Wolf, 1998b). One of the first research that use Dependency/frequency graphs and tables, was published by Weijters and van der Aalst (2001). One can remark that the year 2004 has been a very important year for process mining, since most of the algorithms, tools and methods have been proposed in this year. Although many algorithms, tools and methods have been created, some have also been altered and improved. An example is the Alpha (α) algorithm that was first proposed by van der Aalst et al. (2004). Over the course of years, different variations of the algorithm have been proposed. Some examples are the α^+ (de Medeiros et al., 2004; Wen et al., 2006) and $\alpha^\#$ (Wen et al., 2010) algorithm. Another example are the Heuristic algorithms. Weijters et al. (2006) proposed the HeuristicsMiner algorithm. One year later, Rozinat et al. (2007) came up with the Heuristic Miner. In 2009 Mans et al. proposed the Heuristic mining algorithm. The latest algorithm has been proposed by Li et al. (2011) and is named the Heuristic algorithm.

Table 3.3: Control-flow discovery algorithms, tools and methods

	Discovery
Control-flow	• Concurrency Discovery technique (Cook & Wolf, 1998b) • General DAG (Agrawal et al., 1998) • Inductive approache and stochastic task graphs (Herbst, 2000) • KTail (Cook & Wolf, 1998a) • Markov (Cook & Wolf, 1998a) • RNet (Cook & Wolf, 1998a) • Dependency/frequency graph (Weijters & van der Aalst, 2001; Weijters & van

der Aalst, 2002; van der Aalst et al., 2007)

- Dependency/frequency table (Weijters & van der Aalst, 2001; Weijters & van der Aalst, 2002; van der Aalst et al., 2007)
- α (van der Aalst et al., 2004)
- α^+ (de Medeiros et al., 2004; Wen et al., 2006)
- induceUniqueNodeNo-RepetitionsSAG (workflow mining algorithm) (Herbst & Karagiannis, 2004)
- Instance EPC (van Dongen & van der Aalst, 2004)
- Instance graph (van Dongen & van der Aalst, 2004)
- Meta-model (block-oriented) (Schimm, 2004)
- Mining Terminated States Set (Gaaloul et al., 2004)
- Moore (state-labeled) and Mealy (transition-labeled) state machines (Cook et al., 2004)
  - Discovery algorithm
- ProcessDiscover (Greco et al., 2004; Greco et al., 2006)
- Statistic activity depencency algorithm (Gaaloul et al., 2004)
- Statistic activity frequency algorithm (Gaaloul et al., 2004)
- TP-Graph (Hwang et al., 2004)
- TP-Itemset (Hwang et al., 2004)
- TP-Sequence (Hwang et al., 2004)
- LearnOrderedWorkflow (Silva et al., 2005)
- HeuristicsMiner algorithm (Weijters et al., 2006)
- MineWorkflow (Greco et al., 2006)
- Two-step approach (van der Aalst et al., 2006; van der Aalst et al., 2010)
  - Transition systems
  - Region theory
- Genetic algorithm (de Medeiros et al., 2007)
- Heuristic Miner (Rozinat et al., 2007b)
- Language based synthesis algorithms (Bergenthum et al., 2007)

	○ Finite Basis of Feasible Places ○ Separating Feasible Places • Heuristic mining algorithm (Mans et al., 2009) • $\alpha^{\#}$ (Wen et al., 2010) • Clustering algorithm (Li et al., 2011) • Heuristic algorithm (Li et al., 2011) • Probabilistic workflow mining (Ma et al., 2011) ○ LearnOracles-FromLog ○ LearnWfN-FromOracles

Every algorithm, tool and method is proposed based on certain assumptions. One of the assumptions we can find the most, is that the workflow logs contain perfect information. This means that the log is complete (every possible path is present) and doesn't contain noise (parts of logs can be incorrect, incomplete or refer to exceptions). This is for example the case for the α algorithm, its variants and heuristic algorithms. Another assumption that can be found very often, is that every task or activity has a unique label or name. This is for example the case for the Clustering algorithm, the α^{+} algorithm and Dependency/frequency graphs and tables. Other examples are that the process models are block structured (Clustering algorithm and Heuristic algorithm) and that Petri nets must have a single start and end place (Genetic algorithm). Van Dongen, Alves de Medeiros and Wen (2009) noted that the α algorithm and its variants always make three important assumptions. The first and second assumption are already mentioned (no noise and complete logs). The third assumption is that the process model should be expressed in terms of a Petri net and must not contain certain constructs (sequence, parallelism, choices, loops, non-free-choice, invisible tasks and duplicate tasks). The most frequent assumptions are:

- Event logs contain perfect information (complete logs and no noise);
- Workflow models are sound;
- All activities in a process model have unique labels;
- A certain process language is chosen.

Every algorithm, tool or method has its limitations. Examples of these limitations are noise (α algorithm, Instance EPC and Instance graphs), loops (α algorithm and HeuristicsMiner algorithm) and overfitting (Clustering algorithm and Heuristic algorithm). Van Dongen et al. (2009) made a comparison of Petri net discovery algorithms. The results are shown in table 3.4. The algorithms are compared based on completeness, constructs, abstraction and fitness. Constructs are divided into sequence (seq), parallelism (par), choices (cho), loops (lo), non-free-choice (nfc), invisible tasks (it) and duplicate tasks (dt). The variants of the α algorithm are designed to deal with the limitations of the α algorithm. For example, the α^{+} algorithm was developed to deal with short loops, whereas the $\alpha^{\#}$ algorithm can also deal with invisible tasks. The Heuristics Miner has been developed to deal with noise. It is also capable of dealing with most constructs, like sequence, choice, parallelism, loops, invisible tasks. The algorithm is not able to deal with every aspects of non-free-choice and duplicate tasks. The Genetic algorithm can deal with the same constructs as the Heuristics Miner. It can even deal with all kinds of non-free-choice.

Table 3.4: Comparison of Petri net discovery algorithms (van Dongen et al., 2009)

Algorithm	Complete-ness	Constructs							Abstraction	Fitness
		seq	par	cho	lo	nfc	it	dt		underfitting overfitting
α	DS	+	+	+	+/−	−	−	−	1:1	
α+	DS+	+	+	+	+	−	−	−	1:1	
tsinghua-α	CD	+	+	+	+	−	−	−	1:2	
α++	DS++	+	+	+	+	+	−	−	1:1	
α#	DS+	+	+	+	+	−	+	−	1:0..1	
α*	DS	+	+	+	+/−	−	−	+/−	1..n:1	
Heuristic Miner	ES	+	+	+	+	+/−	+	−	1:0..1	
Genetic Alg.	TS	+	+	+	+	+	+	−	1:0..1	
Dupl. GA	TS	+	+	+	+	+	+	+	1..n:0..1	
LangReg Basis	GC	+	+	+	+	+	−	−	1:1	
LangReg Sep	GC	+	+	+	+	+	−	+	1:1	
LangReg ILP	none	+	+	+	+	+	−	−	1:1	
State Discovery	none	+	+	+	+	+	+	+	1..*:0..*	

3.2.2 Control-flow conformance

Control-flow conformance tries to compare an existing process model with a discovered model from event logs of that same process. The difference with discovery is that conformance deals with a-priori information. With control-flow conformance, an organization can detect, locate and explain possible deviations, and measure the severity of them (van der Aalst, 2011). The output of these finding will be presented in terms of a certain process language.

An overview of control-flow conformance algorithms, tools and methods can be found in table 3.5. When we take a look, we see that first research has been published in 2005. One can also remark that more algorithms, tools and methods have been developed instead of improved. Since 2005 until 2009, different researchers proposed a different algorithm, tool or method every year. The only one that is used in multiple researches, is the Conformance Checker (Rozinat & van der Aalst, 2006; Rozinat et al., 2007; Rozinat & van der Aalst, 2008; Rozinat et al., 2009a).

Table 3.5: Control-flow conformance algorithms, tools and methods

	Conformance
Control-flow	• EMiT (Dustdar et al., 2005) • Instance EPC (Multi Phase mining plug-in) (van Dongen & van der Aalst, 2005) • Instance graph (Multi Phase mining plug-in) (van Dongen & van der Aalst, 2005) • MinSoN (Dustdar et al., 2005) • Transformation Algorithm (Multi Phase mining plug-in) (van Dongen & van der Aalst, 2005) • Conformance Checker (Rozinat & van der Aalst, 2006; Rozinat et al., 2007b; Rozinat & van der Aalst, 2008; Rozinat et al., 2009a) • LTL Checker (Rozinat et al., 2007b) • Genetic Programming technique (Turner et al., 2008)

Most of these algorithms and tools were checked using real life logs. In some cases the logs had to be converted into the MXML format, otherwise the information could not be extracted. This was the case for the Conformance Checker and the LTL Checker. Also, in some cases a selection of logs was made. In case of the conformance Checker and the LTL Checker, not all logs were selected because of the huge amount of event logs. A selection of event logs by machines was made based on four criteria. The first one is that the test process had to be completed. Second, it only had to include the test period. Third, the machines had to belong to the same family. And fourth, it could not be a pilot system, since a pilot system is used for development testing instead of manufacturing qualification. In case of the EMiT and MinSoN tools, assumptions were made that all work cases belong to the same real world process, because for ad-hoc processes it cannot be determined automatically to which real world process they belong. To give an overview of the assumptions:

- When there are too many logs, a selection must be made based on criteria;
- When dealing with real business processes, one should consider clustering.

When looking at the limitations, we can remark that in the case of the EMiT and MinSoN tool the task names in the different cases, which belong to the same task in the real word, had to be same. Otherwise useful process mining was not possible. In case of the Genetic Programming it could only deals with the evolution of graph structures in which the numbers of nodes and their functional behavior were fixed.

3.2.3 Control-flow enhancement

When an organization wants to perform control-flow enhancement, its goal is to extend or improve its existing business process based of the information from event logs of the current process. Just as conformance, enhancement deals with a-priori information. Enhancement can be split up into two types: repair and extension. Repair means modifying the model to give a better representation of the reality. Extension means adding new perspectives to the model (van der Aalst, 2011). Also here, al the output will be presented in terms of a certain process language.

The algorithms, tools and methods concerning control-flow enhancement are given in table 3.6. As we can see, only two tools are mentioned by Dustdar et al. (2005). We can remark that these tools can also be found in table 3.5 concerning control-flow conformance.

Table 3.6: Control-flow enhancement algorithms, tools and methods

	Enhancement
Control-flow	• EMiT (Dustdar et al., 2005) • MinSoN (Dustdar et al., 2005)

The assumption and limitations of the EMiT and MinSoN tool are already described in the part of control-flow conformance.

3.2.4 Organizational discovery

One of the goal of organizational discovery is to create a model of the organizational structure by classifying people based on their roles. Another goal is trying to discover social networks. When people complain about the workload, organizational discovery makes it possible to discover the flow of tasks within a company.

Table 3.7 gives an overview of some algorithms, tools and methods. Here we can see that not so much research had been spend on this type of analysis. One of the oldest tools is MiSoN (van der Aalst & Song, 2004; van der Aalst et al., 2005; van der Aalst et al., 2007). Since the proposal of this tool, it took until 2009 for the proposal of another tool. Mans et al. (2009) proposed the Handover of Work metric and the Social Network Miner.

Table 3.7: Organizational discovery algorithms, tools and methods

	Discovery
Organization	• MiSoN (van der Aalst & Song, 2004; van der Aalst et al., 2005; van der Aalst et al., 2007)

	• Handover of Work metric (Mans et al., 2009) • Social Network Miner (Mans et al., 2009)

The Handover of Work metric and the Social Network Miner were tested using real life processes. To avoid spaghetti-like models, the logs were preprocessed. The limitation in case of the MiSoN tool, is that is only capable of monitoring events that are actually logged.

3.2.5 Organizational conformance

Just as control-flow conformance, organizational conformance tries to align the existing organizational structure with the current structure using information from event logs. An example is to check whether certain tasks that need multiple approvals aren't performed by one person (van der Aalst, 2011).

In table 3.8, some conformance tools and methods are mentioned. We can again see the EMiT and MinSoN tools by Dustdar et al. Other tools and methods are Conformance Testing and the Delta Analysis presented by van der Aalst (2005). One can notice that all these tools were presented in 2005 and that almost no extended research has been done to improve existing tools or propose other tools.

Table 3.8: Organizational conformance algorithms, tools and methods

	Conformance
Organization	• Conformance Testing (van der Aalst, 2005) • Delta Analysis (van der Aalst, 2005) • EMiT (Dustdar et al., 2005) • MinSoN (Dustdar et al., 2005)

The assumptions made in case of the Delta Analysis and Conformance Testing included that events are actually logged by some information systems, that people are not completely controlled by the

system and that they only focused on specific tasks. The limitation of the Delta Analysis is that it is not able to provide quantitative measures for the fit between the prescriptive/ descriptive model and the log. This is where the Conformance Testing comes in. The assumption and limitations of the EMiT and MinSoN tool are already described in the part of control-flow conformance.

3.2.6 Organizational enhancement

The goal of organizational conformance is to extend or improve the existing organizational structure using information from event logs from the current processes. This way organizations will be able to restructure the organization to improve the workflow and decrease the workload.

As we can see in table 3.9, again the tools by Dustdar et al. are mentioned. No other algorithms, tools or methods have been found during this literature study. One can conclude that this type of analysis has not been researched that much.

Table 3.9: Organizational enhancement algorithms, tools and methods

	Enhancement
Organization	• EMiT (Dustdar et al., 2005) • MinSoN (Dustdar et al., 2005)

The assumption and limitations of the EMiT and MinSoN tool are already described in the part of control-flow conformance.

3.2.7 Case discovery

When an organization performs a case discovery analysis, its goal may be trying to create a model of all the originators working on a certain case. The focus of the case perspective is on the properties of cases.

Table 3.10 gives an overview of some algorithms, tools and methods concerning case discovery. As we can see, not much research has been done concerning this analysis. In this table we can see that every two years a new tool or method has been proposed instead of improving an existing one.

Table 3.10: Case discovery algorithms, tools and methods

	Discovery
Case	• Answer Tree (SPSS tool) (van der Aalst et al., 2007) • Dotted chart (Mans et al., 2009) • MinCover (Walicki & Ferreira, 2011)

In the case of the dotted chart, the logs were preprocessed since it was tested using real life logs. If there wasn't any preprocessing, there was a chance ending up with spaghetti-like models. The dotted chart was used because it is able to handle flexible and knowledge intensive business processes. In case of the MinCover tool, the assumption was made that patterns did not contain repeating symbols. The MinCover tool is able to deal with noise and parallelism. The limitations in case of the Answer Tree where that logs were not complete and contained noise, it could only monitor the events that were actually logged and that the system may enforced certain interaction patterns.

3.2.8 Case conformance

When an organization already has an existing model, it can compare it with the current model made from the information in event logs. The organization can for example verify if rules concerning certain amounts of money are followed. An example can be that orders below a certain amount must follow a different path than orders above that certain amount. This way the organization can detect fraud. This has a link with other perspectives.

During this literature study, no research has been found concerning case conformance analyses. It is possible though that case conformance algorithms, tools or methods do exist or are being developed.

3.2.9 Case enhancement

Case enhancement goes a step further than case conformance. Instead of only comparing the current model with the existing model, the organization tries to extend or improve the current model. In the example of checking if people are authorized to do certain tasks, the organization can for example improve the process by creating new guidelines. For example, orders with an amount higher than €25.000 must be authorized by a higher staff. This has also a link with other perspectives.

During this literature study, no research has been found concerning case enhancement analyses. It is possible though that case enhancement algorithms, tools or methods do exist or are being developed.

3.2.10 Time discovery

One of the goals of time discovery is to create a model to check the running times of different kind of activities. Based on this information, organizations can make predictions. Time discovery is only possible if event logs contain timestamps.

As we can see in table 3.11, only one article has been found concerning a time discovery algorithm. This can be explained by the fact that the time perspective is relatively new. In the older research articles, researchers explained that there were only three perspectives: the process perspective, the organizational perspective and the case perspective (Weijters et al., 2006).

Table 3.11: Time discovery algorithms, tools and methods

	Discovery
Time	• Interval algorithm (Pinter & Golani, 2004)

The interval algorithm assumes that there were no directed cycles in the graph, that every label appears at most once in each execution and that was a selection of logged records. The only issue of concern that was discussed during the proposal of the algorithm, was the recall value (the ratio of correctly indentified edges over the total number of edges in the original workflow graph). The algorithm is able to deal with noise.

3.2.11 Time conformance

When an organization performs a time conformance analysis, its goal could be to check whether the utilization of the resources in the current model matches a predefined model. Just as time discovery, time conformance is only possible when event logs contain timestamps.

During this literature study, no research has been found concerning time conformance analyses. It is possible though that time conformance algorithms, tools or methods do exist or are being developed.

3.2.12 Time enhancement

The goal of a time enhancement analysis is to create model using current information from event logs containing timestamps and compare them with an existing model. Based on the new information, the existing model is extended or improved. This way bottlenecks can be discovered and solved.

Table 3.12 gives an overview of a few time enhancement tools and methods. We can see that only two tools and methods are give. The reason why there aren't so many, is that the time perspective is quite new.

Table 3.12: Time enhancement algorithms, tools and methods

	Enhancement

Time	• Basic Log Statistics plug-in (Rozinat et al., 2007b, Rozinat et al., 2009b)
	• CDPMS (Ho et al., 2009)
	  ○ PME
	  ○ DR²M

In both cases, the logs contained real life information and they only focused on a certain part of the logs. In the case of CDPMS, only eleven main processes of three production sites were considered. The logs that were used to test the Basic Log Statistics plug-in were chosen based on four criteria. They had to be completed, only include the test period, belong to the same family and not be part of a pilot system. The plug-in is able to deal with less structured processes.

3.3 Process mining problems

As we can see above, every algorithm and tool has its limitations. In 2004, van der Aalst and Weijters identified the most challenging problems within process mining. Some examples are hidden tasks, mining loops, noise, etc. Although some of them have been solved, many problems remain. Tiwari, Turner and Majeed (2008) investigated papers about process mining that discuss these types of problems. A detailed overview of these papers can be found in table 3.13 in the appendices. Figure 3.3 gives a visual overview of the number of papers compared with the type of problems. Here we can conclude that most papers address the issue of noise, namely twelve papers discuss this problem. Ten papers discuss the problems concerning mining loops and concurrent processes, while the problem of visualizing results is discussed by eight papers. Figure 3.4 gives an overview of the numbers of papers written each year on the problems of process

mining. We can conclude that there has been a rising interest in process mining problems between 2001 and 2004, where it reached its peak. But still a high number of papers were written in 2005.

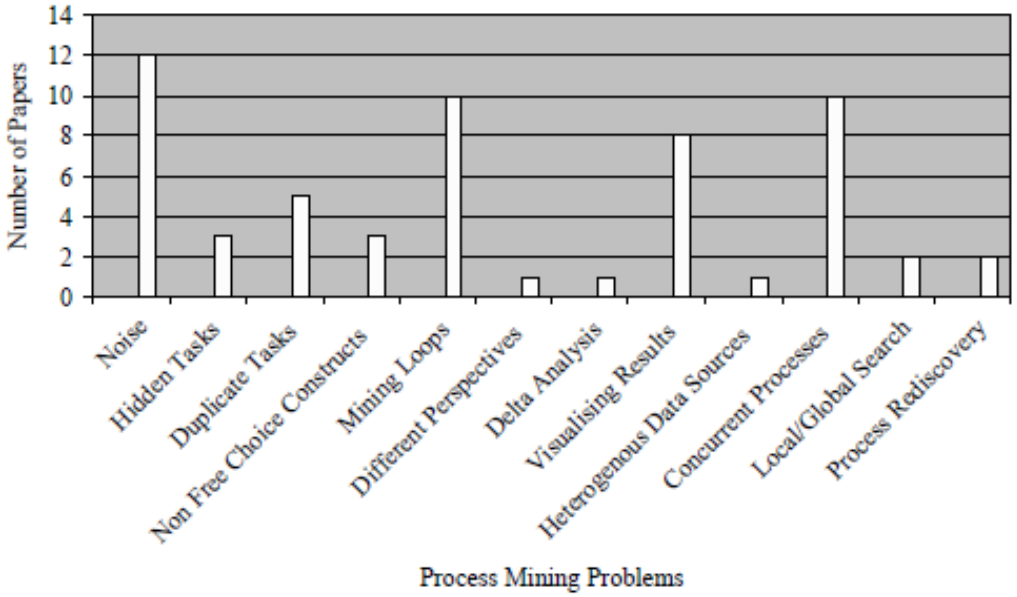

Figure 3.3: Papers dealing with process mining problems (Tiwari et al., 2008)

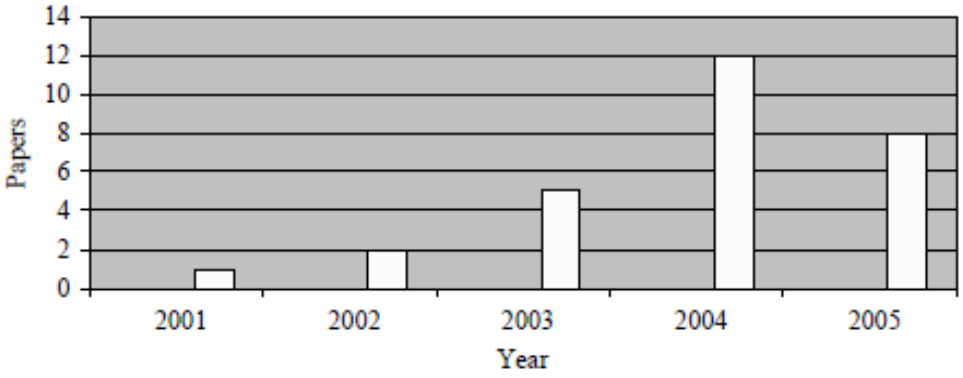

Figure 3.4: Process mining problems (Tiwari et al., 2008)

3.4 Process mining challenges

The IEEE Task force on Process Mining created a list eleven challenging aspects concerning process mining (van der Aalst, 2012). The first challenge is about finding, merging and cleaning event data. It is possible that data is spread out over the entire organization. Therefore it is important to

merge all this information. Another thing is the incompleteness of data. It is common that timestamps are missing. And when timestamps are present, it is possible that they have different levels of granularity. Therefore, the goal is to obtain perfect event logs using better tools and methodologies.

The second challenge is about dealing with complex event logs having diverse characteristics. Event logs may contain so little information that it is difficult to make reliable conclusions whereas other event log contain so much information making it difficult to handle. Also, event logs must handle event logs based on open world assumption. This means that although something did not happen it doesn't mean that it cannot happen.

The third challenge is creating representative benchmarks. Many techniques are offered by many different vendors. There are many differences in functionality and performance, so it is difficult to compare the quality between all these techniques. Some metrics for measuring quality already exists, for example fitness, simplicity, precision and generalization. In the field of data mining, many good benchmarks do exist.

Dealing with concept drift is the fourth challenge. Concept drift means that the situation is changing while it's being analyzed. This is possible due to periodic/seasonal changes or changing conditions. It can be discovered by splitting event logs into smaller logs and analyzing them.

The fifth challenge is improving the representational bias used for process discovery. Process discovery techniques visualize models using a particular language (Petri nets, Workflow nets, BPMN, EPC, etc.). When a languages is chosen, it comes with several implicit assumptions. This limits the search space because processes cannot be discovered when the modeling language cannot represent it. This is called a representational bias and it should be a conscious choice, not a choice driven by the preferred graphical representation.

There should also be a balance between quality criteria. Examples of quality criteria are fitness, simplicity, precision and generalization. This is the sixth challenge. A good fitness means that a model contains most of the behavior seen in the event log. Simplicity means that the best model is also the simplest model. A model that is precise does not allow for too much behavior. When this is not the case, a model is seen as underfitting. This means that the model allow for much more behavior from what is seen in the event log. When a model is overfitting, it does not generalize. This means that a model may explain one event log, but not another one.

Cross-organizational mining is seen as a seventh challenge. In today's world, information is not held within one company but within multiple. Yet, process mining is carried out within a single organization. Cross-organizational mining can be carried out within two settings. The first one is a collaborative setting where different organizations work together. The second one is a setting where organizations do it on their own, but share experiences and knowledge. To do this, new analysis techniques need to be developed.

The eighth challenge is proving operational support. Process mining used to be done using historic data, but nowadays data can be present in real-time. Process mining should not only be done for offline, but also for online operational support. There are three types of operational support activities, namely detect, predict and recommend. Predefined models can be used to detect when a case starts to deviate from its normal path. Predictive models, which can be used as guidance, can be build by using historical data. Based on a predictive model, it is possible to create a recommender system that proposes the most proper action.

Combining process mining with other types of analysis is a ninth challenge. In the field of operations management and data mining, a lot of techniques have been developed. Examples of operations management techniques are queueing models, Markov chains and simulations. Examples of data mining techniques are classification, regression, clustering and pattern discovery. That is why it can be useful to combine process mining with these fields. It can also be useful to combine process mining with visual analytics. The goal of visual analytics is to better understand

large and complex data sets using a combination of automated analysis and interactive visualization.

The tenth challenge is trying to improve the usability for non-experts. End-users need to use process mining results on a daily basis. The challenge lies in creating user-friendly interfaces in a way that it suggests suitable types of analysis and automatically sets parameters.

Improving understandability for non-experts is the eleventh and final challenge. It will not always be easy to understand and interpret process mining results. When one does not understand the output, one can easily make false conclusions. A problem for most current techniques, is that they don't give an indication of the fitness. This is why the trustworthiness of results should be presented.

CHAPTER 4: Data mining and knowledge discovery in databases

Process mining originated from data mining. Whereas process mining tries to discover underlying processes, data mining is all about extracting useful data out of databases using data analysis and discovery algorithms. Data mining is a part of the knowledge discovery in databases (KDD) process. Figure 4.1 gives an overview of the steps in the KDD process.

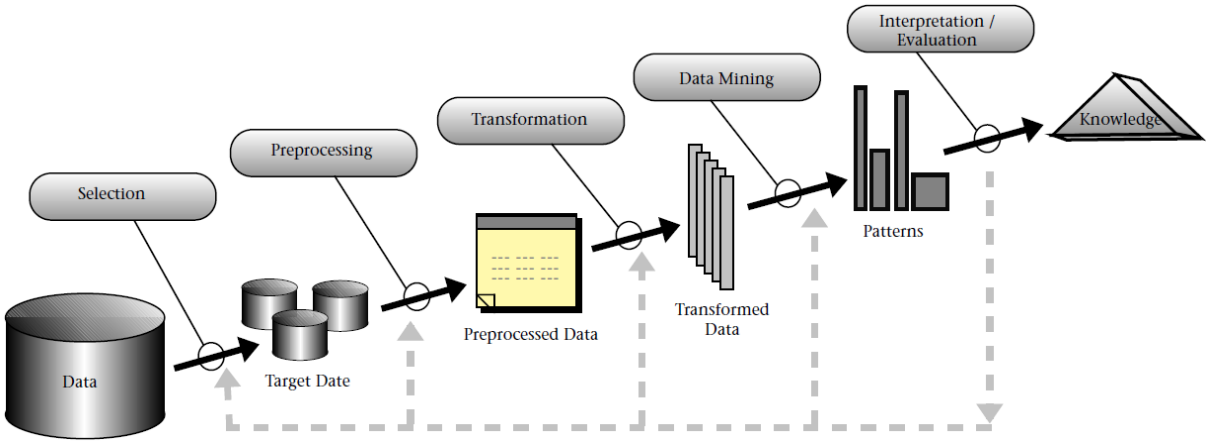

Figure 4.1: Steps of the KDD process (Fayyad, Piatetsky-Shapiro & Smyth, 1996a)

4.1 KDD process models

The framework proposed by Fayyad, Piatetsky-Shapiro & Smyth (1996b) consists of nine steps. The first step is developing and understanding of the application domain. The second step is the creation of a target data set on which the discovery will be performed. The third step consists the cleaning and preprocessing of data. This means the data is for example checked for noise. The fourth step is data reduction and projection. The fifth step is choosing a data mining method (summarization, classification, regression, clustering, etc.) that matches the goal in the first step. The sixth step is choosing the data mining algorithm. The seventh step is the actual data mining itself. The eighth step is interpreting the mined patterns. It is even possible to return to the previous seven steps to make adjustments. The ninth and final step is consolidating the discovered knowledge (Fayyad et al., 1996b; Kurgan & Musilek, 2006).

Table 4.1: Comparison of the major existing KDD models (Kurgan & Musilek, 2006)

Model	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N/A
No of steps	9	5	8	6	6	6
Refs	(Fayyad *et al.*, 1996d)	(Cabena *et al.*, 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios *et al.*, 2000)	N/A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
			2 Problem Specification			
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting	2 Data Understanding	2 Understanding the Data	2 Data Understanding
			4 Domain Knowledge Elicitation			
	3 Data Cleaning and Preprocessing		5 Methodology Identification	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	4 Data Reduction and Projection		6 Data Preprocessing			
	5 Choosing the DM Task					
	6 Choosing the DM Algorithm					
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

More frameworks concerning KDD can be found in the literature. Kurgan and Musilek give an overview of the major existing KDD models (see table 4.1). Next to the model of Fayyad et al., four other models are discussed. The second model was proposed by Cabena, Hadjinian, Stadler, Verhees and Zanasi (1998) and consists out of five steps. The third model was introduced by Anand and Büchner (1998) and consist out of eight steps. The fourth model, named the CRISP-DM (Cross-Industry Standard Process for DM) model consists out of six steps and was first introduced by Shearer (2000). The fifth and final model by Cios, Teresinska, Konieczna, Potocka, and Sharma (2000) consists out of six steps and was influenced by the CRISP-DM model. Kurgan and Musilek (2006) also mentioned several other models in their research paper, but didn't discussed them in detail since they made a less significant impact.

4.2 Analysis of the KDD process models

Kurgan and Musilek compared the five KDD models mentioned above with a generic model. This is shown in table 4.2. The generic model consists out of six steps. The first step is understanding the application domain. The second step is data understanding. Data preparation and identification of DM technology is the third step. The fourth step is data mining. After the mining, the evaluation takes place, which is the fifth step. The sixth and final step is the consolidation and deployment of the gained knowledge.

4.2.1 Step 1: application domain understanding

The first step by Fayyad et al. consists of two parts: learning the goals of the end-user and relevant prior knowledge. Cabena et al. consider the first step as understanding the business problem and defining business objectives which are later redefined into data mining goals. The only difference is that Cabena et al. redefine the problems and objectives into data mining goals. Anand and Büchner see the first step as the identification of human resources and their roles. Anand and Büchner also mention the portioning of the project into smaller tasks. This way it will be easier to solve them using a particular data mining method. Shearer's first step is also the understanding of

Table 4.2: Detailed description of individual steps of the major existing KDD models (Kurgan & Musilek, 2006)

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Steps	STEP 1. Application Domain Understanding	1 Learning goals of the end-user and relevant prior knowledge	1 Understanding the business problem and defining business objectives, which are later redefined into DM goals	1 Identification of human resources and their roles 2 Partitioning of the project into smaller tasks that can be solved using a particular DM method	1 Understanding of business objectives and requirements, which are converted into a DM problem definition	1 Defining project goals, identifying key people, learning current solutions and domain terminology, translation of project goals into DM goals, and selection of DM methods for Step 4
	STEP 2 Data Understanding	2. Selection of a subset of variables and sampling of the data to be used in later steps	2 Identification of internal and external data sources, selection of subset of data relevant to a given DM task. It also includes verifying and improving data quality, such as noise and missing data. Determination of DM methods that will be used in the next step and transformation of the data into analytical model required by selected DM methods	3 Analysis of accessibility and availability of data, selection of relevant attributes and a storage model 4 Elicitation of the project domain knowledge	2 Identification of data quality problems, data exploration, and selection of interesting data subsets	2 Collecting the data, verification of data completeness, redundancy, missing values, plausibility, and usefulness of the data with respect to the DM goals

Table 4.2: Detailed description of individual steps of the major existing KDD models (continued) (Kurgan & Musilek, 2006)

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Steps	STEP 3 Data Preparation and Identification of DM Technology	3 Preprocessing of noise, outliers, missing values, etc, and accounting for time sequence information 4 Selection of useful attributes by dimension reduction and transformation, development of invariant data representation 5 Goals from Step 1 are matched with a particular DM method, i.e. classification, regression, etc. 6 Selection of particular data model(s), method(s), and method's parameters		5 Selection of the most appropriate DM method, or a combination of DM methods 6 Preprocessing of the data, including removal of outliers, dealing with missing and noisy data, dimensionality reduction, data quantization, transformation and coding, and resolution of heterogeneity issues	3 Preparation of the final dataset, which will be fed into DM tool(s), and includes data and attribute selection, cleaning, construction of new attributes, and data transformations	3 Preprocessing via sampling, correlation and significance tests, cleaning, feature selection and extraction, derivation of new attributes, and data summarization. The end result is a data set that meets specific input requirements for the selected DM methods
	STEP 4 Data Mining	7 Generation of knowledge (patterns) from data, for example classification rules, regression model, etc.	3 Application of the selected DM methods to the prepared data	7 Automated pattern discovery from the preprocessed data	4 Calibration and application of DM methods to the prepared data	4 Application of the selected DM methods to the prepared data, and testing of the generated knowledge

Table 4.2: Detailed description of individual steps of the major existing KDD models (continued) (Kurgan & Musilek, 2006)

Model	Generic	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*
Steps	STEP 5 Evaluation	8 Interpretation of the model(s) based on visualization of the model(s) and the data based on the model(s)	4 Interpretation and analysis of DM results; usually visualization technique(s) are used	8 Filtering out trivial and obsolete patterns, validation and visualization of the discovered knowledge	5 Evaluation of the generated knowledge from the business perspective	5 Interpretation of the results, assessing impact, novelty and interestingness of the discovered knowledge. Revisiting the process to identify which alternative actions could have been taken to improve the results
	STEP 6 Knowledge Consolidation and Deployment	9 Incorporation of the discovered knowledge into a final system, creation of documentation and reports, checking and resolving potential conflicts with previously held knowledge	5 Presentation of the generated knowledge in a business-oriented way, formulation of how the knowledge can be exploited, and incorporation of the knowledge into organization's systems		6 Presentation of the discovered knowledge in a customer-oriented way. Performing deployment, monitoring, maintenance, and writing final report	6 Deployment of the discovered knowledge. Creation of a plan to monitor the implementation of the discovered knowledge, documenting the project, extending the application area from the current to other possible domains

business objectives and requirements. They are also converted into a data mining problem definition. Cios et al. splits the first step up into five parts. The first part is defining the project goals. This is the same as Fayyad et al., Cabena et al. and Shearer. The second part is identifying the key people, which is the same as Anand and Büchner. Learning current solutions and domain terminology, which is the third part, can be linked to relevant prior knowledge by Fayyad et al. The fourth part contains the translation of project goals into data mining goals, which is the same for Cabena et al. and Shearer. The last part is the selection of data mining methods. This can be linked with Anand and Büchner.

4.2.2 Step 2: data understanding

Fayyad et al. divide this step into two parts: selection of a subset of variables and sampling of the data. The second step by Cabena et al. is divided into smaller parts. There is the identification of internal and external data sources. There is also the selection of a subset of data relevant to a given data mining task. This can be linked with Fayyad et al. They also propose verifying and improving data quality. Also the determination of data mining methods is mentioned. This is what Anand and Büchner and Cios et al. proposed in the previous step. The last part is the transformation of the data into an analytical model. Anand and Büchner mention the analysis of accessibility and availability of data, which can be linked with the internal and external data sources proposed by Cabena et al. The next thing they mention is the selection of relevant attributes and a storage model. They also mention the elicitation of the project domain knowledge. Shearer proposes the identification of data quality problems, which is also mentioned by Cabena et al. Next to this, data exploration is also mentioned. This was also mentioned by Cabena et al. and Anand & Büchner. The last thing is the selection of interesting data subjects. This was also proposed by Fayyad et al., Cabena et al. and Anand and Büchner. Cios et al. first mention the collection of data, which is mentioned before by Cabena et al., Shearer and Anand and Büchner. The last thing Cios et al. mention the verification of data completeness, redundancy, plausibility and usefulness. This is also mentioned by Cabena et al. and Shearer.

4.2.3 Step 3: data preparation and identification of data mining technology

When we look at the third step of the generic model, we can see that four of the nine steps from the framework by Fayyad et al. are placed here. First they mention the preprocessing of noise, outliers, missing values, etc. This was mentioned by Cios et al., Cabena et al. and Shearer in the previous step. Another part is the accounting for time sequence information. The selection of useful attributes is also proposed in this step. Also a data mining method is chosen that matches the goals as described in the first step. The final part of this step is the selection of particular data models, methods and parameters. When we look at the model of Anand and Büchner, we can see that they also mention the selection of the most appropriate data mining method and the preprocessing of the data. Shearer mentions the preparation of the final data set. This includes data and attribute selection, cleaning, construction of new attributes, and data transformations. Cios et al. propose to do the preprocessing via sampling, correlation and significance tests, cleaning, feature selection and extraction, derivation, and data summarization.

4.3.4 Step 4: data mining

Fayyad et al. describe this step as the generation of knowledge from data, which is the same for Anand and Büchner. Cabena et al., Shearer and Cios et al. describe this as the calibration application of the selected data mining methods. Cios et al. also mention the testing of the generated knowledge.

4.3.5 Step 5: evaluation

All the models describe this step as the interpretation or validation of the results (models and data) using visualization techniques. Shearer even explains to do the evaluation from the business perspective. Anand and Büchner also mention filtering out trivial and obsolete patterns. To complement this, Cios et al. propose revisiting the process to identify which alternative actions could be taken to improve the results.

4.3.6 Step 6: knowledge consolidation and deployment

Fayyad et al. describe this step as the incorporation of the discovered knowledge intro a final system. They also mention the creation of documentation and reports. As a final remark, they propose to check and resolve potential conflicts with previously held knowledge. Cabena et al. also mention the documentation and incorporation of the knowledge. Next to this, they also propose a presentation of the knowledge in a business-oriented way. Shearer on the contrary, proposes to do the presentation in a customer-oriented way. Shearer also mentions the monitoring and maintenance of the knowledge. Al these proposals, except the presentations, were also mentioned by Cios et al. As a final remark, they propose to extend the application area from the current to other possible domains.

CHAPTER 5: Process mining framework

Rozinat et al. (2007a) state that although "process mining reached a certain level of maturity … a common framework to evaluate process mining results is still lacking" (p.1). This comes with a fact that researchers can't compare the performance of different process mining algorithms. A similar framework can be composed for algorithms, tools and methods concerning the different types and perspectives of process mining. Based on the frameworks introduced in KDD, we will try to make a similar framework for the fourteen different process mining analyses. Table 5.1 gives a summary of all the steps of the KDD models and its sub steps. As we can see, the summary of the KDD models has no logic sequence since different sub steps can be found in multiple steps.

Table 5.1: Summary of the KDD models

STEP 1	**Application domain understanding**
	Identification of human resources and their roles
	Learning relevant prior knowledge
	Learning goals of the end user
	Understand business problems and define business objectives
	Translate objectives into data mining goals
	Partitioning of the project into smaller tasks
	Selection of data mining methods
STEP 2	**Data understanding**
	Selection of a subset of variables and data
	Selection of relevant attributes and a storage model
	Sampling of data
	Identification of internal and external data sources
	Verifying and improving data quality
	Determination of data mining methods
	Elicitation of the project domain knowledge

	Transformation of the data into an analytical model
STEP 3	**Data preparation and identification of data mining technology**
	Preprocessing of data
	Accounting for time sequence information
	Selection of useful attributes and data
	Matching goals with a particular data mining method
	Selection of particular data models, methods and method's parameters
STEP 4	**Data mining**
	Calibration and application of the selected data mining methods
	Generation of knowledge from data
	Testing generated knowledge
STEP 5	**Evaluation**
	Interpretation of the results
	Evaluation of the generated knowledge from the business perspective
	Filtering out trivial and obsolete patterns
	Revisiting the process for possible improvements
STEP 6	**Knowledge consolidation and deployment**
	Incorporation of the discovered knowledge
	Creation of documents and reports
	Check and resolve potential conflicts with previously held knowledge
	Presentation in a business-oriented way
	Presentation in a customer-oriented way
	Monitoring the knowledge
	Maintaining the knowledge
	Extending to other possible domains

5.1 Linking KDD steps to process mining

With the help of two users of process mining, the different steps and sub steps of the KDD process models have been linked to process mining. The two users are both PhD students. One uses process mining only in an academic setting, while the other also applies process mining for companies that want a better insight of their activities and processes. In the following subsections, the different sub steps from the KDD process models will be discussed if they can be used for the process mining framework.

5.1.1 Step 1: application domain understanding

The identification of the human resources and their roles can be used when performing process mining. When trying to discover social networks, one can check if tasks are truly performed by people appointed to do them.

Learning relevant prior knowledge can also be useful when organizations want to try process mining on their own, instead of consulting companies specialized in process mining. Relevant prior knowledge can also mean checking if some predetermined models of activities and tasks exist within the company.

Learning goals of the end user and understanding business problems can help to determine to define business objectives. Keeping these objectives in mind, one or more of the fourteen different analyses can be chosen. This can be seen as translating objectives into process mining goals. After translating objectives into process mining goals, one or more of the fourteen process mining analyses can be chosen.

When there are too many objectives defined or many analyses are chosen, it can be possible to split up the project into smaller parts. An example can be to first create a process model, and later on perform a social network analysis on that process model.

5.1.2 Step 2: data understanding

Before one can begin process mining, one must look where data about processes can be found. Therefore it is important to first identify internal and external data sources. Internal data can be found within an organization, while external data is located outside of the organization. When too many data has been found, it can be useful to select a subset with enough data to perform process mining. Also, it can be useful to sample the data and perform an analysis on the sample. After that, it can be possible to perform an analysis on the data as a whole, to check if the results of the sample are conform with the data as a whole. When the dataset is chosen, it is also important, especially when doing an analysis concerning the case perspective, to select relevant attributes. After all these steps it is suggested to verify and improve data quality by filtering out irrelevant and adding relevant data and attributes.

When a process mining analysis is chosen, it is important to chose one or more tools that are able to perform that specific type of analysis. Examples of such tools can be found in the third chapter. It is also important to transform the data into a model or language that be interpret by the chosen process mining tool. An example is that data sometimes has to be converted into the MXML format. The elicitation of the project domain knowledge has been discussed in the previous step.

5.1.3 Step 3: data preparation and identification of data mining technology

The preprocessing of data can also be used in process mining. This can also be seen as the transformation of data, for example into the MXML format. Accounting for time sequence information is also important. The presence of this information allows to order activities into a certain sequence. This can be classified under selection, verification and improving of data. For example, data without the presence of time information should be filtered out, since it won't be able to use them to determine the sequence of activities. The selection of useful attributes and data has already been discussed in the previous step. As mentioned before, the goals and objectives defined in the first step should be matched with a particular process mining method.

First, the type of analysis should be chosen. Finally, the tool(s) to perform that analysis should be determined. This was also discussed in the first and second step.

5.1.4 Step 4: data mining

For process mining, all the sub steps from the fourth step are applicable. The first sub step is the calibration and application of the selected process mining tools that were chosen based on the choice of the process mining analysis. The information from the event logs is used to create process models, social networks, etc. In other words, knowledge is generated. After the knowledge is generated, it should be tested. For example, when only a sample of the data is used, one can try to perform the same analysis on the data as a whole to check if the results are representative.

5.1.5 Step 5: evaluation

When the results are tested, one has to interpret these. Since the process mining expert mostly does not anything about the internal affairs of the organization, it could be useful to get help from someone familiar with the process within the organization, so that no mistakes can be made during the interpretation. This means that instead of only evaluating it in a statistical way, it should also be evaluated from a business perspective.

When evaluating the results, one can perhaps filter out trivial and obsolete patterns. This means that there should be a feedback to the fourth and even to the second step. In general, the process can be revisited for possible improvements. This means that there is a feedback possible to all the previous steps.

5.1.6 Step 6: knowledge consolidation and deployment

When linking all these sub steps to process mining, we can conclude that all these sub steps are applicable. After the evaluation of the generated knowledge, it should be incorporated into the

organization. All the steps and results should also be documented in reports, so the organization can check them later when necessary. When some documentation and knowledge is already present in an organization, it is important to check and resolve potential conflicts with the new discovered knowledge. It is also recommended to present the knowledge to the concerned parties. From one of the interviews appears that the monitoring and maintenance of knowledge is quite important, but most of the organizations don't really apply this. Afterwards, when all the steps for a certain process are finished, one can also check if it is possible to extend process mining to other possible domains within the organization.

5.2 Process diagnostics

Bozkaya, Gabriels and van der Werf (2009) proposed a process diagnostics methodology (see figure 5.1) which consists of five phases. In the first phase, the event log is prepared. This means that data is collected from the information systems. Since data can be found in different forms, it is important to preprocess the log. Preprocessing takes place in several steps. The first step is to select the best notion of a case, whereas the second step is to identify all the activities and their events.

The second phase is the inspection of the log. This means that statistics about the log are gathered. Examples are the number of cases and roles, the total number of events, the average number of events per case, etc. Using these statistics, better insights are obtained about the size of the process and event log. Also using these statistics, the event log is filtered to remove incomplete cases. These are cases that were started before a certain time or that are still not finished.

A control-flow analysis is considered as the third phase. As mentioned in the third chapter, the control-flow perspective tries to generate a model how the actual process within an organization looks like. When the organization already has a predetermined model, it is possible to do a conformance check whether the current model fits or not.

The fourth phase is a performance analysis, which is a combination of the case and time perspective. This analysis tries to uncover bottlenecks in the process. This is done by using a dotted chart analysis, which compares cases and their throughput times. After this, the log is replayed on the process model, which calculates bottlenecks and throughput times of individual activities and the process itself.

The role analysis is the fifth phase. This matches the organizational perspective. As mentioned in the third chapter, it tries to expose who performs what activities and who are working together. The first step is to create a role-activity matrix, where the rows represent the roles and the columns represent each event of the event log. In the second step, the roles are analyzed. Two types of roles can be distinguished, namely specialist and generalists. Specialists are roles that only execute a few activities but very frequently, while generalists are roles that execute many different activities in the process. As a third step, a social network analysis is performed, which reveals relations between roles.

After these five phases, the results have to be transferred. To gain insight in processes of the organization, it is important to discuss the outcomes with the organization. Using this knowledge, the organization is able to adjust the business processes where necessary.

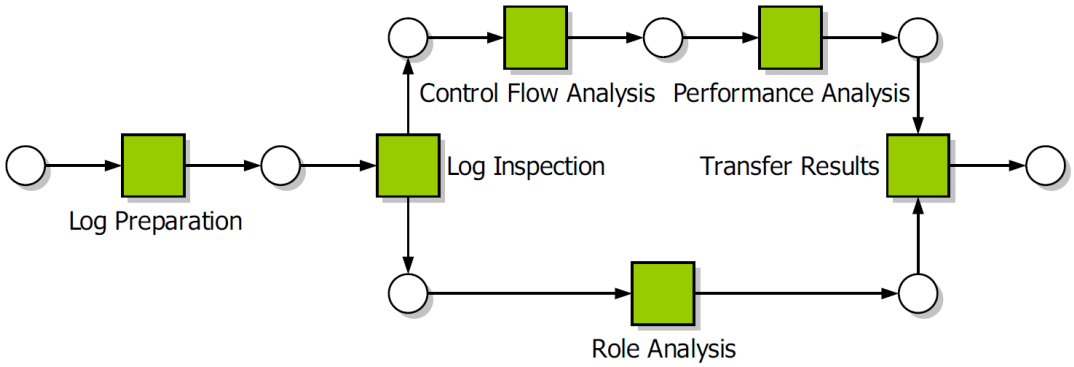

Figure 5.1: The phases of the process diagnostics methodology (Bozkaya et al., 2009)

5.3 Development of a process mining framework

Based on the KDD process models, information from the interviews and the process diagnostics methodology by Bozkaya et al. (2009), a process mining framework can be developed. Figure 5.2 gives an overview of all the big steps. The first step is the goal definition. Searching and preparing the data is the second step. The third step is the creation of event logs and the identification of process mining technology. Process mining is the fourth step. Evaluation of the generated knowledge is the fifth step. The sixth and final step is knowledge consolidation and deployment. These steps can be divided into smaller sub steps and will be explained in the following sections. A complete overview of the process mining framework can be found in table 5.8.

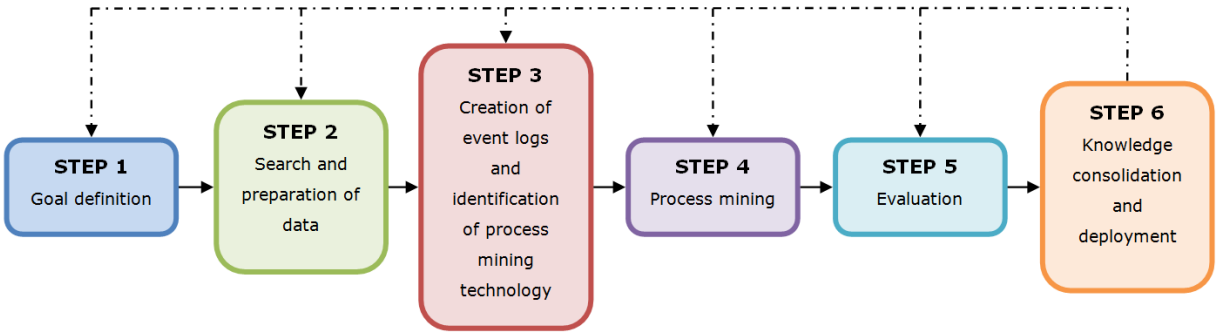

Figure 5.2: Process mining flowchart

5.3.1 Step 1: goal definition

The first sub step is the identification of human resources and their roles. As stated before, it can be useful to ask information about the roles so one can check if activities are truly performed by people appointed to do them. This can be useful when analyzing the organizational perspective. The second sub step is learning relevant prior knowledge. We can distinguish two types: knowledge about the organization itself and knowledge about process mining. Knowledge about the organization can be seen as predetermined models, guidelines, etc. Knowledge about process mining is useful when an organization that is not familiar with process mining wants to conduct an analysis.

After these sub steps, the third and the fourth sub steps are to understand the goal of the end user and the business problems. Based on these problems and goals, objectives can be defined that serve as the basis for the process mining analysis. It is important to translate these business objectives into process mining goals, which is the fifth sub step. When too many goals are found, it can be useful to partition the project into smaller tasks. This is the sixth step. As a seventh and final sub step, the goals have to be matched with one or more of the fourteen process mining analyses that are described in the third chapter.

Table 5.2: Goal definition

STEP 1	**Goal definition**
	Identification of human resources and their roles
	Learning relevant prior knowledge
	Learning goals of the end user
	Understand business problems and define business objectives
	Translate objectives into process mining goals
	Partitioning of the project into smaller tasks
	Matching goals with a process mining analysis

5.3.2 Step 2: search and preparation of data

When all the objectives are defined, the next big step is to search and prepare the data for the analysis. The first sub step is to identify internal and external data sources. Internal data can be found within the organization's own information systems, whereas external data can be found outside the organization, for example partners that are involved in the process. After the identification, the second sub step is to select a subset of data. It is important to select the data with the most useful information. Out of this data, the most relevant attributes and variables have to be selected. This is for example very important when analyzing the case perspective. This is considered as the third sub step. When there is still too much data left after the selection, it is

possible to sample the data. An analysis can be performed on the sample and later on the data as a whole to check for consistency. This is the fourth and final sub step for the second main step.

Table 5.3: Search and preparation of data

STEP 2	**Search and preparation of data**
	Identification of internal and external data sources
	Selection of a subset of data
	Selection of relevant attributes and variables
	Sampling of data

5.3.3 Step 3: creation of event logs and identification of process mining technology

To perform process mining on a data set, an event log should be created. In some cases it is necessary to first preprocess the data before it can be turned into an event log. An example is that data mostly must be converted in the MXML format. Preprocessing can be considered as the first sub step. The second sub step is the creation of the event logs. After the creation, the event logs should be checked if the quality can be improved by adding relevant or removing irrelevant data. This means that there is a feedback to the second main step. Finally, one (or more) process mining tool has to be selected. The tool must be capable of performing the type of analysis that was determined in the first step. Examples of tools per analysis can be found in the third chapter.

Table 5.4: Creation of event logs and identification of process mining technology

STEP 3	**Creation of event logs and identification of process mining technology**
	Preprocessing of data
	Creation of event logs
	Verifying and improving data quality
	Selection of process mining tools

5.3.4 Step 4: process mining

When the process mining tools are chosen, they first need to be calibrated when possible and then applied to the generated event logs. This generated knowledge, which can be seen as the second sub step. As a final sub step, the generated knowledge can be tested. This is the case when one wants to check if the results of process mining performed on the total data is consistent with the results of process mining performed on a sample of that data.

Table 5.5: Process mining

STEP 4	**Process mining**
	Calibration and application of the selected process mining tools
	Generation of knowledge from data
	Testing generated knowledge

5.3.5 Step 5: evaluation

After the generation of knowledge, it should be evaluated. The first sub step therefore is the interpretation of the results. It is recommended that the interpretation is done in cooperation with people familiar with the organization. This way, misunderstandings can be avoided. Instead of only evaluating it in a statistical way, it should also been seen from a business perspective. Also here the cooperation with someone familiar with the organization is recommended.

The results should be checked if patterns occur. One can consider to filter out these patterns if they are trivial and obsolete. If these patterns are frequent or even devious, it can be interesting to analyze them. In all the cases, the process mining step has to be redone. This means there is a feedback to the fourth main step. As a final sub step, it can be possible to revisit all the main and sub steps to if improvements are possible for better results. This means that there is a feedback to all the previous main steps.

Table 5.6: Evaluation

STEP 5	**Evaluation**
	Interpretation of results
	Evaluation of the generated knowledge from the business perspective
	Filtering out trivial and obsolete patterns
	Revisiting the process for possible improvements

5.3.6 Step 6: knowledge consolidation and deployment

When all the results are evaluated, first they should be presented to all the parties that are a part of the analyzed process. As a second sub step, all the results should be documented in reports in case someone requests them. During the documentation, one should check and resolve if there are potential conflicts with previously held information. This way, no redundant information about the process is present. The fourth sub step is to incorporate the discovered knowledge. When it is incorporated, it is important to monitor and maintain that knowledge. As a final sub step, it can be possible to extend the analyses to other domains. This means that all will start again from the beginning.

Table 5.7: Knowledge consolidation and deployment

STEP 6	**Knowledge consolidation and deployment**
	Presentation of the discovered knowledge
	Creation of documents and reports
	Check and resolve potential conflicts with previously held knowledge
	Incorporation of discovered knowledge
	Monitoring the knowledge
	Maintaining the knowledge
	Extending to other possible domains

Table 5.8: Process mining framework

STEP 1	**Goal definition**
	Identification of human resources and their roles
	Learning relevant prior knowledge
	Learning goals of the end user
	Understand business problems and define business objectives
	Translate objectives into process mining goals
	Partitioning of the project into smaller tasks
	Matching goals with a process mining analysis
STEP 2	**Search and preparation of data**
	Identification of internal and external data sources
	Selection of a subset of data
	Selection of relevant attributes and variables
	Sampling of data
STEP 3	**Creation of event logs and identification of process mining technology**
	Preprocessing of data
	Creation of event logs
	Verifying and improving data quality
	Selection of process mining tools
STEP 4	**Process mining**
	Calibration and application of the selected process mining tools
	Generation of knowledge from data
	Testing generated knowledge
STEP 5	**Evaluation**
	Interpretation of results
	Evaluation of the generated knowledge from the business perspective
	Filtering out trivial and obsolete patterns
	Revisiting the process for possible improvements

STEP 6	**Knowledge consolidation and deployment**
	Presentation of the discovered knowledge
	Creation of documents and reports
	Check and resolve potential conflicts with previously held knowledge
	Incorporation of discovered knowledge
	Monitoring the knowledge
	Maintaining the knowledge
	Extending to other possible domains

CHAPTER 6: Discussion

The process mining framework is based on the most known KDD process models. This means that there are resemblances but also some differences. The main resemblance is that the models are almost the same when it comes to the main steps. As we can see, the first step of the KDD process models and the process mining model is completely the same. The sub steps of this step share the same sequence.

The second step of both models involves searching and selecting data. The four sub steps as proposed in the process mining framework can also be found in the KDD process models. We can notice that the determination of a data mining method can again be found in this step. This because some authors of the KDD models place this sub step in a different step. Also the verification and improvement of the data quality can be found in the second step of the KDD models, and in the third step of the process mining model. The transformation of the data in the second step of the KDD models can also be found in the third step of the process mining model in the form of creation of event logs.

The similarity between the third step of the KDD models and the process mining model is the identification of the mining technology. Two similar sub steps can be found in both models. The first one is the preprocessing of data, the second one is the selection of the mining tools. Again, selecting a mining method can be found in this step of the KDD models. As mentioned before, this is done in the first step of the process mining model. Also the selection of useful attributes and data is done in the second step op the process mining model.

The fourth, fifth and sixth steps of the KDD process models are all similar with the process mining model. For the fourth and fifth step, all sub steps have the same sequence. For the sixth step, the only difference is that the incorporation and presentation of the knowledge are in a different order. In the KDD models, one first incorporates the knowledge and then presents it, while in the process mining model it is vice versa.

CHAPTER 7: Conclusions and recommendations

7.1 Conclusions

Based on the previous chapters, we are able to formulate conclusions that provide an answer for the research question and the sub questions as proposed in the first chapter.

Process mining has two dimensions. The first dimension is the process mining type. Three different types of process mining can be distinguished: discovery, conformance and enhancement. Discovery means that there is no a-priori model. Conformance means that there is an a-priori model and is used to check if the actual model conforms to the predefined model. Enhancement also means that there is an a-priori model, but here the goal is to enrich the model using information from event logs. The other dimension is the process mining perspective. There are four different perspectives: the control-flow perspective, the organizational perspective, the case perspective and the time perspective.

In the field of knowledge discovery in databases, different frameworks have been proposed. The most widely used ones were proposed by Fayyad et al., Cabena et al., Anand and Büchner, Shearer and Cios et al. When comparing all these frameworks, we can see that they have six big steps in common. The first step is understanding the domain. The second step is understanding the data. The preparation of data and the identification of data mining technology is the third step. Data mining is considered as the fourth step. The evaluation of the results is the fifth step. The sixth and final step is the consolidation and deployment of the gained knowledge.

Based on interviews with process mining users, the different steps and sub steps of the KDD process models have been linked with process mining. The conclusion can be made that most of the KDD sub steps were also relevant for process mining. However, the order of some sub steps for process mining were sometimes different from the KDD sub steps. With the additional use of a process diagnostic methodology proposed by Bozkaya et al. (2009), a process mining framework

has been created. The framework consists of six main steps. The first step is defining goals and objectives and has seven sub steps. The second step is searching and preparing data and contains four sub steps. The creation of event logs and identifying the process mining technology is the third step and can be divided into four sub steps. The actual process mining itself is the fourth step, which in its turn contains three sub steps. The fifth step is the evaluation of the generated results and knowledge. This step has four sub steps. The consolidation and deployment of that knowledge is the sixth and final step, and contains seven sub steps.

7.2 Recommendations

The framework (see table 5.8) has been created based on interviews and KDD process models. It however has not been tested using artificial or real-life data. For future research, it could be useful to test the proposed flowchart while conducting the fourteen process mining analyses.

REFERENCES

Agrawal, R., Gunopulos, D., & Leymann, F. (1998). Mining Process Models from Workflow Logs [Electronic version]. *Proceedigns of the Sixth International Conference on Extending Database Technology*, 469-483.

Anand, S., & Büchner, A.G. (1998). *Decision Support Using Data Mining*. London: Financial Time Management.

Bergenthum, R., Desel, J., Lorenz, R., & Mauser, S. (2007). Process Mining Based on Regions of Languages [Electronic version]. *Proceedings of the 5th international conference on Business process management*.

Bibiano, L.H., Mayol, E., & Paster, J.A. (2007). *Role and Importance of Business Processes in the Implementation of CRM Systems*. Requested on December 8, 2011, via http://alarcos.inf-cr.uclm.es/pnis/articulos/pnis-07-bibiano-raibpcrm.pdf.

Bozkaya, M., Gabriels, J., & van der Werf, J.M. (2009). Process Diagnostics: a Method Based on Process Mining [Electronic version]. *Proceedings of International Conference on Information, Process, and Knowledge Management*.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering Data Mining: From Concepts to Implementation*. New Jersey: Prentice Hall.

Champy, J. (1995). *Reengineering Management: Mandate for New Leadership*. HarperCollins e-books.

Cios, K.J., Teresinska, A., Konieczna, S., Potocka, J., & Sharma, S. (2000). A knowledge discovery approach to diagnosing myocardial perfusion [Electronic version]. *IEEE Engineering in Medicine and Biology Magazine*, *19:4*, 17-25.

Cook, J.E., & Wolf, A.L. (1998a). Discovering Models of Software Processes from Event-Based Data [Electronic version]. *ACM Transactions on Software Engineering and Methodology*, *7:3*, 215-249.

Cook, J.E., & Wolf, A.L. (1998b). Event-Based Detection of Concurrency [Electronic version]. *Proceedings of the Sixth International Symposium on the Foundations of Software Engineering*, 35-45.

Cook, J.E., Du, Z., Liu, C., & Wolf, A.L. (2004). Discovering models of behavior for concurrent workflows [Electronic version]. *Computers in Industry*, *53*, 297-319.

de Medeiros, A.K.A., van Dongen, B.F., van der Aalst, W.M.P., & Weijters, A.J.M.M. (2004). Process Mining for Ubiquitous Mobile Systems: An Overview and a Concrete Algorithm [Electronic version]. *Lecture Notes in Computer Science*, *3272*, 151-165.

de Medeiros, A.K.A., Weijters, A.J.M.M., & van der Aalst, W.M.P. (2007). Genetic process mining: an experimental evaluation [Electronic version]. *Data Mining and Knowledge Discovery*, *14:2*, 245-304.

Dustdar, S., Hoffmann, T., & van der Aalst, W. (2005). Mining of ad-hoc business processes with TeamLog [Electronic version]. *Data & Knowledge Engineering*, *55*, 129-158.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From Data Mining to Knowledge Discovery in Databases [Electronic version]. *AI Magazine*, *17:3*, 37-54.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). Knowledge Discovery and Data Mining: Towards a Unifying Framework [Electronic version]. *Proc. 2nd Int. Conf on Knowledge Discovery and Data Mining*, 82-88.

Gaaloul, W., Bhiri, S., & Godart, C. (2004). Discovering Workflow Transactional Behavior from Event-Based Log [Electronic version]. *Lecture Notes in Computer Science*, *3290*, 3-18.

Greco, G., Guzzo, A., Pontieri, L., & Saccà, D. (2004). Mining Expressive Process Models by Clustering Workflow Traces [Electronic version]. *Lecture Notes in Artificial Intelligence*, *3056*, 52-62.

Greco, G., Guzzo, A., Pontieri, L., & Saccà, D. (2006). Discovering Expressive Process Models by Clustering Log Traces [Electronic version]. *IEEE Transactions on Knowledge and Data Engineering*, *18:8*, 1010-1027.

Herbst, J. (2000). A Machine Learning Approach to Workflow Management [Electronic version]. *Lecture Notes in Computer Science*, *1810*, 183-194.

Herbst, J., & Karagiannis, D. (2004). Workflow mining with InWoLvE [Electronic version]. *Computers in Industry*, *53*, 245-264.

Ho, G.T.S., Lau, H.C.W., Kwok, S.K., Lee, C.K.M., & Ho, W. (2009). Development of a co-operative distributed process mining system for quality assurance [Electronic version]. *International Journal of Production Research*, *47:4*, 883-918.

Hwang, S.Y., Wei, C.P., & Yang, W.S. (2004). Discovery of temporal patterns from process instances [Electronic version]. *Computers in Industry*, *53*, 345-364.

Internet Growth Statistics. (2011). Requested on December 8, 2011, via http://www.internetworldstats.com/emarketing.htm.

Kurgan, L.A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models [Electronic version]. *The Knowledge Engineering Review*, *21:1*, 1-24.

Li, C., Reichert, M., & Wombacher, A. (2011). Mining business process variants: Challenges, scenarios, algorithms [Electronic version]. *Data & Knowledge Engineering*, *70*, 409-434.

Ma, H., Tang, Y., & Wu, L. (2011). Incremental Mining of Processes With Loops [Electronic version]. *International Journal on Artificial Intelligence Tools*, *20:1*, 221-235.

Mans, R.S., Schonenberg, M.H., Song, M., van der Aalst, W.M.P., & Bakker, P.J.M. (2009). Application of Process Mining in Healthcare – A Case Study in a Dutch Hospital [Electronic version]. *Communications in Computer and Information Science*, *25:4*, 425-438.

Moore's Law. (2011). Requested on November 22, 2011, via http://www.investopedia.com/terms/m/mooreslaw.asp#axzz1eS6TesVM.

Moore's Law Graph. (2011). Requested on November 22, 2011, via http://www.intel.com/pressroom/kits/events/moores_law_40th/index.htm.

Peppard, J., & Ward, J. (2005). Unlocking Sustained Business Value from IT Investments [Electronic version]. *California Management Review*, *48:1*, 52-70.

Pinter, S.S., & Golani, M. (2004). Discovering workflow models from activities' lifespans [Electronic version]. *Computers in Industry*, *53*, 283-296.

ProM. (2009). Requested on November 28, 2011, via http://www.processmining.org/prom/start.

Rozinat, A., & van der Aalst, W.M.P. (2006). Conformance Testing: Measuring the Fit and Appropriateness of Event Logs and Process Models [Electronic version]. *Lecture Notes in Computer Science*, *3812*, 163-176.

Rozinat, A., & van der Aalst, W.M.P. (2008). Conformance checking of processes based on monitoring real behavior [Electronic version]. *Information Systems*, *33*, 64-95.

Rozinat, A., Alves de Medeiros, A.K., Günther, C.W., Weijters, A.J.M.M., & van der Aalst, W.M.P. (2007a). Towards an Evaluation Framework for Process Mining Algorithms [Electronic version]. *BPM Center Report BPM-07-06*, 1-20.

Rozinat, A., de Jong, I.S.M., Günther, C.W., & van der Aalst, W.M.P. (2007b). Process Mining of Test Processes: A Case Study [Electronic version]. *BETA Working Paper Series*, *WP 220*, 1-36.

Rozinat, A., de Jong, I.S.M., Günther, C.W., & van der Aalst, W.M.P. (2009a). Conformance Analysis of ASML's Test Process [Electronic version]. *Proceedings of the Second International Workshop on Governance, Risk and Compliance*, 1-15.

Rozinat, A., de Jong, I.S.M., Günther, C.W., & van der Aalst, W.M.P. (2009b). Process Mining Applied to the Test Process of Wafer Steppers in ASML [Electronic version]. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, *39:4*, 474-479.

Rundle, M. (2010). *Why Apple Succeeds & Others Fail*. Requested on December 8, 2011, via http://flyosity.com/apple/why-apple-succeeds-others-fail.php.

Schimm, G. (2004). Mining exact models of concurrent workflows [Electronic version]. *Computers in Industry*, *53*, 265-281.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5:4*, 13-22.

Silva, R., Zhang, J., & Shanahan, J.G. (2005). Probabilistic Workflow Mining [Electronic version]. *Proc. of 11th Intl. Conf. Knowledge Discovery in Data Mining*, 275-284.

Summer, M. (2007). *Enterprise Resource Planning* (Hamers, R., van Kuijk, M., & van Lumig, A., Dutch). Pearson Education Benelux: Amsterdam. (2005).

Timeline. (2011). Requested on December 8, 2011, via http://www.facebook.com/press/info.php?timeline.

Tiwari, A., Turner, C.J., & Majeed, B. (2008). A review of business process mining: state-of-the-art and future trends [Electronic version]. *Business Process Management Journal*, *14:1*, 5-22.

Turner, C.J., Tiwari, A., & Mehnen, J. (2008). A Genetic Programming Approach to Business Process Mining [Electronic version]. *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, 1307-1314.

van der Aalst, W.M.P. (2005). Business alignment: using process mining as a tool for Delta analysis and conformance testing [Electronic version]. *Requirements Engineering*, *10:3*, 198-211.

van der Aalst, W.M.P. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Berlin: Springer.

van der Aalst et al. (2012). Process Mining Manifesto [Electronic version]. *Lecture Notes in Business Information Processing*, *99*, 169-194.

van der Aalst, W.M.P., & Song, M. (2004). Mining Social Networks: Uncovering Interaction Patterns in Business Processes [Electronic version]. *Lecture Notes in Computer Science*, *3080*, 244-260.

van der Aalst, W.M.P., & Weijters, A.J.M.M. (2004). Process mining: a research agenda [Electronic version]. *Computers in Industry*, *53*, 231-244.

van der Aalst, W.M.P., Reijters, H.A., & Song, M. (2005). Discovering Social Networks from Event Logs [Electronic version]. *Computer Supported Cooperative Work*, *14*, 549-593.

van der Aalst, W.M.P., Reijers, H.A., Weijters, A.J.M.M., van Dongen, B.F., Alves de Medeiros, A.K., Song, M., & Verbeek, H.M.W. (2007). Business Process Mining: An Industrial Application [Electronic version]. *Information Systems*, *32:5*, 713-732.

van der Aalst, W.M.P., Rubin, V., van Dongen, B.F., Kindler, E., & Günther, C.W. (2006). Process Mining: A Two-Step Approach using Transition Systems and Regions [Electronic version]. *BPM Center Report BPM-06-30*, *BPM Center*, *13*.

van der Aalst, W.M.P., Rubin, V., Verbeek, H.M.W., van Dongen, B.F., Kindler, E., Günther, C.W. (2010). Process mining: a two-step approach to balance between underfitting and overfitting [Electronic version]. *Softw Syst Model*, *9*, 87-111.

van der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow Mining: Discovering Process Models from Event Logs [Electronic version]. *IEEE Transactions on Knowledge and Data Engineering*, *16:9*, 1128-1142.

van Dongen, B.F., & van der Aalst, W.M.P. (2004). Multi-phase Process Mining: Building Instance Graphs [Electronic version]. *Lecture Notes in Computer Science*, *3288*, 362-376.

van Dongen, B.F., & van der Aalst, W.M.P. (2005). Multi-Phase Process Mining: Aggregating Instance Graphs into EPCs and Petri Nets [Electronic version]. *Proc. Int'l Workshop Applications of Petri Nets to Coordination, Workflow and Business Process Management*.

van Dongen, B.F., Alves de Medeiros, A.K., & Wen, L. (2009). Process Mining: Overview and Outlook of Petri Net Discovery Algorithms [Electronic version]. *Lecture Notes in Computer Science*, *5460*, 225-242.

Walicki, M., & Ferreira, D.R. (2011). Sequence partitioning for process mining with unlabeled event logs [Electronic version]. *Data & Knowledge Engineering*, *70*, 821-841.

Weber, P. (2009). *A Framework for the comparison of Process Mining algorithms*. Requested on December 12, 2011, via http://www.cs.bham.ac.uk/~pxw869/papers/MSc/msc_project.pdf.

Weijters, A.J.M.M., & van der Aalst, W.M.P. (2001). Process Mining: Discovering Workflow Models from Event-Based Data [Electronic version]. *Proceedings of the 13th Belgium-Netherlands Conference on Artificial Intelligence*, 283-290.

Weijters, A.J.M.M., & van der Aalst, W.M.P. (2002). Workflow Mining: Discovering Workflow Models from Event-Based Data [Electronic version]. *Proceedings of the ECAI Workshop on Knowledge Discovery and Spatial Data*, 78-84.

Weijters, A.J.M.M., van der Aalst, W.M.P., & Alves de Medeiros, A.K. (2006). Process Mining with the HeuristicsMiner Algorithm [Electronic version]. *BETA publicatie: working papers*, *166*, 1-34.

Wen, L., Wang, J., & Sun, J. (2006). Detecting Implicit Dependencies Between Tasks from Event Logs [Electronic version]. *Lecture Notes in Computer Science*, *3841*, 591-603.

Wen, L., Wang, J., van der Aalst, W.M.P., Huang, B., & Sun, J. (2010). Mining process models with prime invisible tasks [Electronic version]. *Data & Knowledge Engineering*, *69*, 999-1021.

What is Hadoop?. (2011) Requested on December 12, 2011, via http://www.cloudera.com/what-is-hadoop/.

Winter, R. (2008). *Why Are Data Warehouses Growing So Fast? An Update on the Drivers of Data Warehouse Growth*. Requested on December 12, 2011, via http://www.b-eye-network.com/view/7188.

APPENDICES

A Tables

Table 2.1: Fragment of an event log (van der Aalst, 2011)

Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	Register request	Pete	50	...
	35654424	31-12-2010:10.06	Examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	Check ticket	Mike	100	...
	35654426	06-01-2011:11.18	Decide	Sara	200	...
	35654427	07-01-2011:14.24	Reject request	Pete	200	...
2	35654483	30-12-2010:11.32	Register request	Mike	50	...
	35654485	30-12-2010:12.12	Check ticket	Mike	100	...
	35654487	30-12-2010:14.16	Examine casually	Pete	400	...
	35654488	05-01-2011:11.22	Decide	Sara	200	...
	35654489	08-01-2011:12.05	Pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	Register request	Pete	50	...
	35654522	30-12-2010:15.06	Examine casually	Mike	400	...
	35654524	30-12-2010:16.34	Check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	Decide	Sara	200	...
	35654526	06-01-2011:12.18	Reinitiate request	Sara	200	...
	35654527	06-01-2011:13.06	Examine thoroughly	Sean	400	...
	35654530	08-01-2011:11.43	Check ticket	Pete	100	...
	35654531	09-01-2011:09.55	Decide	Sara	200	...
	35654533	15-01-2011:10.45	Pay compensation	Ellen	200	...
4	35654641	06-01-2011:15.02	Register request	Pete	50	...
	35654643	07-01-2011:12.06	Check ticket	Mike	100	...
	35654644	08-01-2011:14.43	Examine thoroughly	Sean	400	...
	35654645	09-01-2011:12.02	Decide	Sara	200	...
	35654647	12-01-2011:15.44	Reject request	Ellen	200	...
5	35654711	06-01-2011:09.02	Register request	Ellen	50	...
	35654712	07-01-2011:10.16	Examine casually	Mike	400	...
	35654714	08-01-2011:11.22	Check ticket	Pete	100	...
	35654715	10-01-2011:13.28	Decide	Sara	200	...
	35654716	11-01-2011:16.18	Reinitiate request	Sara	200	...
	35654718	14-01-2011:14.33	Check ticket	Ellen	100	...
	35654719	16-01-2011:15.50	Examine casually	Mike	400	...
	35654720	19-01-2011:11.18	Decide	Sara	200	...
	35654721	20-01-2011:12.48	Reinitiate request	Sara	200	...
	35654722	21-01-2011:09.06	Examine casually	Sue	400	...
	35654724	21-01-2011:11.34	Check ticket	Pete	100	...
	35654725	23-01-2011:13.12	Decide	Sara	200	...
	35654726	24-01-2011:14.56	Reject request	Mike	200	...
6	35654871	06-01-2011:15.02	Register request	Mike	50	...
	35654873	06-01-2011:16.06	Examine casually	Ellen	400	...
	35654874	07-01-2011:16.22	Check ticket	Mike	100	...
	35654875	07-01-2011:16.52	Decide	Sara	200	...
	35654877	16-01-2011:11.47	Pay compensation	Mike	200	...
...	...	...	...	...	...	...

Table 3.2: Type/Perspective-matrix of authors

		TYPES OF PROCESS MINING		
		Discovery	**Conformance**	**Enhancement**
PERSPECTIVE	**Control-flow**	• Agrawal, Gunopulos & Leymann (1998) • Bergenthum, Desel, Lorenz & Mauser (2007) • Cook & Wolf (1998a) • Cook & Wolf (1998b) • Cook, Du, Liu & Wolf (2004) • de Medeiros, van Dongen, van der Aalst & Weijters (2004) • de Medeiros, Weijters & van der Aalst (2007) • Gaaloul, Bhiri & Godart (2004) • Greco, Guzzo, Pontieri & Saccà (2004) • Greco, Guzzo, Pontieri & Saccà (2006) • Herbst (2000) • Herbst & Karagiannis (2004) • Hwang, Wei & Yang (2004) • Li, Reichtert & Wombacher (2011)	• Dustdar, Hoffmann & van der Aalst (2005) • Rozinat & van der Aalst (2006) • Rozinat & van der Aalst (2008) • Rozinat, de Jong, Günther & van der Aalst (2007b) • Rozinat, de Jong, Günther & van der Aalst (2009a) • Turner, Tiwari & Mehnen (2008) • van Dongen & van der Aalst (2005)	• Dustdar, Hoffmann & van der Aalst (2005)

		• Ma, Tang & Wu (2011) • Mans, Schonenberg, Song, van der Aalst & Bakker (2009) • Rozinat, de Jong, Günther & van der Aalst (2007b) • Schimm (2004) • Silva, Zhang & Shanahan (2005) • van der Aalst, Weijters & Maruster (2004) • van der Aalst, Rubin, van Dongen, Kindler & Günther (2006) • van der Aalst et al. (2007) • van der Aalst et al. (2010) • van Dongen & van der Aalst (2004) • Weijters & van der Aalst (2001) • Weijters & van der Aalst (2002) • Weijters, van der Aalst & Alves de Medeiros (2006) • Wen, Wang & Sun (2006) • Wen, Wang, van der		

		Aalst, Huang & Sun (2010)		
	Organization	• Mans, Schonenberg, Song, van der Aalst & Bakker (2009) • van der Aalst & Song (2004) • van der Aalst, Reijers & Song (2005) • van der Aalst et al. (2007)	• Dustdar, Hoffmann & van der Aalst (2005) • van der Aalst (2005)	• Dustdar, Hoffmann & van der Aalst (2005)
	Case	• Mans, Schonenberg, Song, van der Aalst & Bakker (2009) • van der Aalst et al. (2007) • Walicki & Ferreira (2011)		
	Time	• Pinter & Golani (2004)		• Ho, Lau, Kwok, Lee & Ho (2009) • Rozinat, de Jong, Günther & van der Aalst (2007b) • Rozinat, de Jong, Günther & van der Aalst (2009b)

Table 3.13: Papers dealing with process mining problems (Tiwari et al., 2008)

	Noise	Hidden tasks	Duplicate tasks	Non-free choice constructs	Mining loops	Different perspectives	Delta analysis	Visualising results	Heterogenous data sources	Concurrent processes	Local global search	Process rediscovery
van der Aalst (2005)							X					
Greco *et al.* (2004)											X	
Cook and Wolf (1998b)										X		
van der Aalst *et al.* (2002)				X								X
Dongen *et al.* (2005b)								X				
Weijters and van der Aalst (2003)	X				X							X
Golani and Pinter (2003)										X		
Cook and Wolf (1998a)	X							X				
Alves de Medeiros *et al.* (2004c)					X			X		X		
Zhang *et al.* (2003)										X		
Alves de Medeiros *et al.* (2004b)	X				X							
Gaaloul and Godart (2005)	X									X		
Alves de Medeiros *et al.* (2003)					X							
Dongen and van der Aalst (2004b)	X											
Dongen and van der Aalst (2004a)								X				
Alves de Medeiros *et al.* (2004a)					X							
Dongen and van der Aalst (2005b)			X									
Agrawal *et al.* (1998)	X		X									
Alves de Medeiros *et al.* (2005)	X	X		X								

(continued)

Table 3.13: Papers dealing with process mining problems (continued) (Tiwari et al., 2008)

	Noise	Hidden tasks	Duplicate tasks	Non-free choice constructs	Mining loops	Different perspectives	Delta analysis	Visualising results	Heterogenous data sources	Concurrent processes	Local global search	Process rediscovery
Weijters and van der Aalst (2001)	X				X					X		
Schimm (2003)					X							
Dongen and van der Aalst (2005a)									X			
Hammori *et al.* (2004)								X		X		
Cook *et al.* (2004)	X											
Hwang (2002)	X											
van der Aalst *et al.* (2005a)	X	X	X	X	X			X		X	X	
van der Aalst and Alves de Medeiros (2005)	X	X	X		X							
Herbst and Karagiannis (2004)			X		X			X				
Dustdar *et al.* (2004)								X				
van der Aalst and Song (2004)						X						
Cook and Wolf, 1998b)										X		
Schimm (2004)										X		

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Developing a process mining workflow based on the KDD framework

Richting: **Master of Management-Management Information Systems**
Jaar: **2012**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Geuns, Raf

Datum: **30/05/2012**